EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

Valerio Basile
University of Turin
valerio.basile@unito.it

Maria Di Maro
University of Naples "Federico II"
maria.dimaro2@unina.it

Danilo Croce
University of Rome "Tor Vergata"
croce@info.uniroma2.it

Lucia C. Passaro
University of Pisa
lucia.passaro@fileli.unipi.it

1 Introduction

The Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA) is the biennial initiative aimed at promoting the development of language and speech technologies for the Italian language. EVALITA is promoted by the Italian Association of Computational Linguistics (AILC)[1] and it is endorsed by the Italian Association for Artificial Intelligence (AIxIA)[2] and the Italian Association for Speech Sciences (AISV)[3].

EVALITA provides a shared framework where different systems and approaches can be scientifically evaluated and compared with each other with respect to a large variety of tasks, suggested and organized by the Italian research community. The proposed tasks represent scientific challenges where methods, resources, and systems can be tested against shared benchmarks representing linguistic open issues or real world applications, possibly in a multilingual and/or multi-modal perspective. The collected data sets provide big opportunities for scientists to explore old and new problems concerning NLP in Italian as well as to develop solutions and to discuss the NLP-related issues within the community. Some tasks are traditionally present in the evaluation campaign, while others are completely new.

This paper introduces the tasks proposed at EVALITA 2020 and provides an overview to the participants and systems whose descriptions and obtained results are reported in these Proceedings[4]. The EVALITA 2020 edition, held online on December 17^{th} due to the COVID-19 pandemic, counts 14 different tasks. In particular, the selected tasks are grouped in five research areas (tracks) according to their objective and characteristics, namely (i) *Affect, Hate, and Stance*, (ii) *Creativity and Style*, (iii) *New Challenges in Long-standing Tasks*, (iv) *Semantics and Multimodality*, (v) *Time and Diachrony*.

This edition was highly participated, with 51 groups whose participants have affiliation in 14 countries. Although EVALITA is generally promoted and targeted to the Italian research community, this edition saw an international participation, also thanks to the fact that several Italian researchers working in different countries contributed to the organization of the tasks or participated in them as authors.

This overview is organized as follows: in Section 2 a brief description of the tasks belonging to the various areas is reported. Section 3 discusses the participation to the workshop referred to several aspects, from the research area, to the affiliation of authors. Section 4 describes the criteria used to assign the best system across tasks award, made by an ad-hoc committee starting from the suggestions of task organizers and reviewers. Finally, section 5 points out on both the obtained results and on the future of the workshop.

[1] http://www.ai-lc.it
[2] http://www.aixia.it
[3] http://www.aisv.it
[4] The presentations of these works are publicly available at https://vimeo.com/showcase/evalita2020 . All videos are also grouped according to different tasks at https://vimeo.com/user125537954/albums

2 EVALITA 2020 Tracks and tasks

In the 2020 edition of EVALITA, 14 different tasks were proposed, peer-reviewed and accepted. Data were produced by the task organizers and made available to the participants. For the future availability of this data we are going to release them on GitHub[5], in accordance to the terms and conditions of the respective data sources. Such a repository will also reference alternative repositories managed by the task organizers. The tasks of EVALITA 2020 are grouped according to the following tracks:

Affect, Hate, and Stance

AMI - Automatic Misogyny Identification (Fersini et al., 2020). This shared task is aimed at automatically identifying misogynous content in Twitter for the Italian language. In particular, the AMI challenge is focused on: (1) recognizing misogynous and aggressive messages and (2) discriminating misogynistic contents from the non-misogynistic ones, while guaranteeing the fairness of the model.

ATE_ABSITA - Aspect Term Extraction and Aspect-Based Sentiment Analysis (De Mattei et al., 2020b). A task on Aspect Term Extraction (ATE) and Aspect-Based Sentiment Analysis (ABSA). The task is approached as a cascade of three subtasks: Aspect Term Extraction (ATE), Aspect-based Sentiment Analysis (ABSA) and Sentiment Analysis (SA).

HaSpeeDe - Hate Speech Detection (Sanguinetti et al., 2020). A rerun of the shared task on hate speech detection at the message level on Italian social media texts proposed for the first time in 2018 for the EVALITA evaluation campaign. The main task is a binary hate speech detection task, one in-domain and one out-of-domain. On the same data provided for the main task, the topics of stereotypes in communication and nominal utterances are investigated by of two pilot tasks.

SardiStance - Stance Detection (Cignarella et al., 2020). The goal of the task is to detect the stance of the author towards the target "Sardines movement" in Italian tweets. Two subtasks model (A) Textual Stance Detection and (B) Contextual Stance Detection. Both the subtasks consist on a three-class (in favour, against, neutral) classification problem based on textual information only (A) or on the text provided with additional information about the author and the post of the tweet.

Creativity and Style

CHANGE-IT - Style Transfer (De Mattei et al., 2020a). The first natural language generation task for Italian. Change-IT focuses on style transfer performed on the headlines of two Italian newspapers at opposite ends of the political spectrum. Specifically, the goal is to "translate" the headlines from a style to another.

TAG-it - Topic, Age and Gender Prediction (Cimino et al., 2020). TAG-IT is a profiling task for Italian. It is a follow-up of the GxG task organised in the context of EVALITA 2018. The task is aimed at profiling along with three dimensions (Gender, Age, and Topic). Authors propose several subtasks where participants are asked to predict one or more classes starting from the others.

Semantics and Multimodality

CONcreTEXT - Concreteness in Context (Gregori et al., 2020). The task focuses on automatic assignment of concreteness values to words in context for the Italian and English language. Participants are required to develop systems able to rate the concreteness of a target word in a sentence on a scale from 1 (for fully abstract) to 5 (for maximally concrete).

DANKMEMES - Multimodal Artefacts Recognition (Miliani et al., 2020). The first multimodal task for Italian. The goal of the task is to deal with Italian memes considering textual and visual cues together. Providing a corpus of memes on the 2019 Italian Government Crisis, DANKMEMES features three subtasks: A) Meme Detection, B) Hate Speech Identification, and C) Event Clustering.

[5] https://github.com/evalita2020

Ghigliottin-AI - Evaluating Artificial Players for the Language Game "La Ghigliottina" (Basile et al., 2020b). The task challenges researchers to develop a system able to defeat human players at the language game "La Ghigliottina", which represents one of the most followed and appreciated quiz games in Italy.

PRELEARN - Prerequisite Relation Learning (Alzetta et al., 2020). The task is devoted to automatically inferring prerequisite relations from educational texts. The task consists in classifying prerequisite relations between pairs of concepts distinguishing between prerequisite pairs and non-prerequisite pairs.

Time and Diachrony

DaDoEval - Dating Documents (Menini et al., 2020). The task focuses on assigning a temporal span to a document, by recognising when a document was issued. A first one coarse-grained classification subtask, participants are asked to provide a document with a class encoding the historical period (5 classes). The second Fine-grained classification task requires to attribute documents with a temporal slice of 5 years.

DIACR-Ita - Diachronic Lexical Semantics (Basile et al., 2020a). The first task on automatic detection of lexical and semantic shift for Italian. The task challenges participants to develop systems that can automatically detect if a given word has changed its meaning over time, given contextual information from corpora.

New Challenges in Long-standing Tasks

AcCompl-it- Acceptability & Complexity evaluation (Brunato et al., 2020). The task is aimed at developing and evaluating methods to classify Italian sentences according to Acceptability and Complexity. Given a set of sentences, two independent subtasks focus on predicting their acceptability and complexity rate.

KIPoS - Part-of-speech Tagging on Spoken Language (Bosco et al., 2020). The first task on Part of Speech tagging of spoken language held at EVALITA. Benefiting from the KIParla corpus, a resource of transcribed spoken Italian, the task provides three evaluation exercises focused on formal versus informal spoken texts.

3 Participation

EVALITA 2020 attracted the interest of a large number of researchers from academia and industry, for a total of 51 teams composed of about 130 individuals participating in one or more of the 14 proposed tasks. After the evaluation period, 58 system descriptions were submitted (reported in these proceedings), i.e., a 70% percentage increase with respect to the previous EVALITA edition (Caselli et al., 2018).

Moreover, task organizers allowed participants to submit more than one system result (called runs), for a total of 240 submitted runs. Table 1 shows the different tracks and tasks along with the number of participating teams and submitted runs. The data reported in the table is based on information provided by the task organizers at the end of the evaluation process. Such data represents an overestimation with respect to the systems described in the proceedings. The trends are similar, but there are differences due to groups participating in more than a task, and groups that have not produced a system report.

Differently from the previous EVALITA editions, the organizers were discouraged from distinguishing the submissions between unconstrained and constrained runs[6]. The rationale for this decision is that the recent spread and extensive use of pre-trained word embedding representations, especially as a strategy to initialize Neural Network architectures, challenges this distinction at its very heart. Participation was quite imbalanced across different tracks and tasks, as reported in Figure 1: each rectangle represents a task whose size reflects the number of participants, while the color indicated the corresponding track.

[6]A system is considered *constrained* when using the provided training data only; on the contrary, it is considered *unconstrained* when using additional material to augment the training dataset or to acquire additional resources.

TRACK	TASK	TEAMS	RUNS
Affect, Hate, and Stance	AMI	8	31
	ATE_ABSITA	3	8
	HaSpeeDe	14	27
	SardiStance	12	36
Creativity and Style	CHANGE-IT	0	0
	TAG-it	3	20
New Challenges in Long-standing Tasks	AcCompl-it	2	6
	KIPoS	3	14
Semantics and Multimodality	CONcreTEXT	4	15
	DANKMEMES	5	15
	Ghigliottin-AI	2	2
	PRELEARN	3	14
Time and Diachrony	DaDoEval	2	16
	DIACR-Ita	9	36

Table 1: Number of participating teams and number of runs organized by track and task. The data reported is an overestimation with respect to the systems described in the proceedings (e.g. teams participating in more than a task are counted according to the number of tasks they participated in).

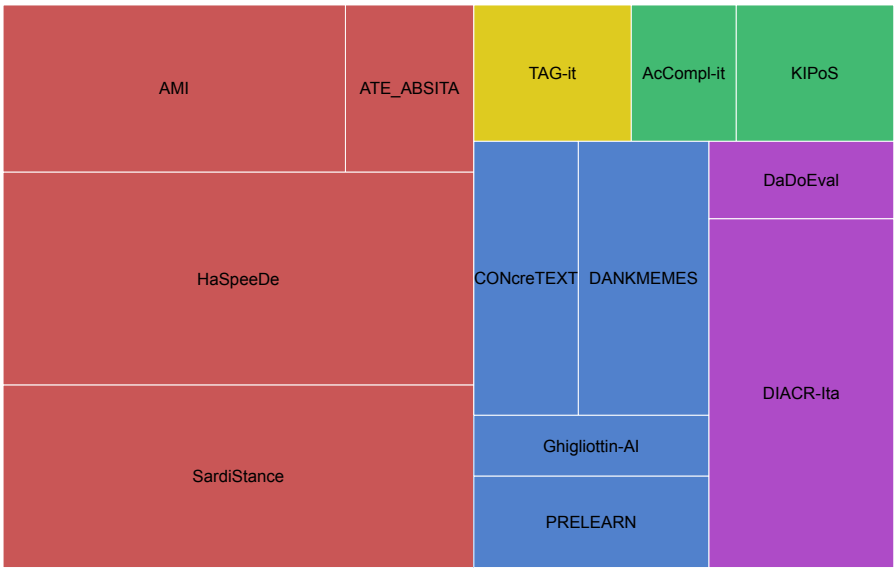

Figure 1: Number of participating teams organized by track (color) and task. The red color is adopted for the track "*Affect, Hate, and Stance*", the yellow color for "*Creativity and Style*", green for "*New Challenges in Long-standing Tasks*", blue for "*Semantics and Multimodality*" and purple for "*Time and Diachrony*".

In line with the previous editions of EVALITA, the track "*Affect, Creativity and Style*" covers about half of the total in terms of participating teams. On the one hand, this demonstrates the well-known interest of the NLP community for Social Media platforms and user-generated content. On the other hand, we report a better balance with respect to the 2018 edition, where about 80% of the teams participated in similar tracks ("*Affect, Creativity and style*" and "*Hate Speech*", which have been merged in this edition). Another significant number of teams participated to the "*Semantics and Multimodality*" and "*Time and Diachrony*" tracks, while the other tracks where less participated. Unfortunately, no team participated to the *CHANGE-IT* task, mainly due to the complexity of the task.

In addition to being widely participated, the over 180 proceedings authors, including both participants and task organizers, have affiliation in 18 countries, with the 64% from Italy and the 36% of participants from Institutions and companies abroad. The group of the 59 task organizers have affiliations in 6 countries (90% from Italy while 10% from Institutions and companies abroad). The gender distribution is highly balanced, with 30 females and males.

4 Award: Best System Across Tasks

In line with the previous edition, we confirmed the award to the best system across-task. The award was introduced with the aim of fostering student participation to the evaluation campaign and to the workshop. EVALITA received sponsorship funding from from Amazon Science, Bnova s.r.l., CELI s.r.l., the European Language Resources Association (ELRA) and Google Research.

A committee of 5 members was asked to choose the best system across tasks. Four of the five members come from academia while the last one is from industry. The composition of the committee is balanced with respect to the level of seniority as well as for their academic background (computer science-oriented vs. humanities-oriented). In order to select a short list of candidates, the task organizers were invited to propose up to two candidate systems participating to their tasks (not necessarily top ranking). The committee was provided with the list of candidate systems and the criteria for eligibility, based on:

- *novelty* with respect to the state of the art;
- *originality*, in terms of identification of new linguistic resources, identification of linguistically motivated features, and implementation of a theoretical framework grounded in linguistics;
- *critical insight*, paving the way to future challenges (deep error analysis, discussion on the limits of the proposed system, discussion of the inherent challenges of the task);
- *technical soundness* and *methodological rigor*.

We collected 10 system nominations from the organizers of 7 tasks from across all tracks. The candidate systems are authored by 20 authors, among whom 12 are students, either at the master's or PhD level. The award recipient(s) will be announced during the final EVALITA workshop, during the plenary session, held online.

5 Final Remarks

A record number of 14 tasks were organized at EVALITA 2020: some of them were revivals of tasks in the past editions (such as *Hate Speech Detection* or *Part-of-Speech Tagging*), while others were completely new (such as the ones involving *Meme Processing* or *Stance Detection*), and were greeted with great enthusiasm by the NLP community.

In this edition, the topics of Affect and Semantics were confirmed as two of the most interesting and thriving ones, both in the number of organized tasks and actual participants. In any case, almost all tasks involved the analysis of written texts. In fact, although the KIPoS task considered transcriptions of spoken Italian utterances, no speech related tasks was proposed.

Anyways, tasks concerning new problems and modalities have been proposed, such as the analysis of memes, and two tasks oriented to the problem of time and diachrony. Moreover, this edition saw an increase in tasks related to creativity and style, despite the fact that one one such tasks, namely CHANGE-it, had no participation, probably due to its complexity and the lack of specific resources for the task in the Italian community. Another task that received a rather low number of submission due to its complexity is GhigliottinAI. Despite being a rather simple word-correlation problem by itself, it required complex modelling of language and semantics to beat the challenge. A very interesting innovation for this task was the evaluation framework, based on APIs, via a Remote Evaluation Server (RES). In general, the most participated tasks have been those by which the linguistic problem could be modelled as a direct classification or regression task.

The competition attracted a record number of participating teams from academia and industry, for a total of 51 teams and more than 180 authors with affiliations in 18 countries. Hopefully, this means that EVALITA is becoming more and more popular also with foreign contributors, and it is becoming an international workshop. First of all, this success confirms the beneficial impact of the organization of the evaluation period based on non-overlapping windows (adopted from EVALITA 2018) in order to help those who want to participate in more than one task. Moreover, we speculate that the technological advancements and ease of use of existing open-source libraries for machine learning and natural language processing improved the accessibility to the tasks, even for master students. In fact, we noticed an

increase in the participation of students, that contributed with state-of-the-art solutions to the tasks. We can argue that the spread of frameworks such as PyTorch and Keras, together with pre-trained, off-the-shelf language models, lowered the set-up costs to deal with complex NLP tasks. In general, we noticed that most of the best systems are based on neural approaches. Among them, BERT or similar Transformer-based architectures achieved the best results: more specifically, at least in 11 out of 14 tasks best results (in at least one sub-task) were obtained by neural architectures based on or combined with Transformers.

We are confident that the positive trends observed in this edition, concerning the participation and the proliferation of tasks, has not yet reached a plateau. It would be auspicabile, among other aspects, to see more tasks involving challenging settings such as, for example, multi-modal or multi-lingual analysis involving Italian, in future EVALITA 2020 editions. Several areas represent fertile ground to organize future tasks, such as domain adaptation (which was considered in previous editions of EVALITA), or few-shot learning to support the definition of robust systems in challenging low-resource settings. Finally, we believe in the importance of defining more structured tasks involving real applications to challenge the Italian community, e.g., Question Answering or Dialogue Agents.

Acknowledgments

We would like to thank our sponsors Amazon Science[7], Bnova[8], CELI[9], European Language Resources Association (ELRA)[10] and Google Research[11] for their support to the virtual event and to the best-system across task award.

Moreover, we gratefully acknowledge the members of the AILC board for their trust and support, our EVALITA advisor Nicole Novielli and all the chairs of the 2018 edition, who helped us during the organization process of EVALITA 2020. In addition, we sincerely thank the Best System across Tasks committee for providing their expertise and experience.

Finally, we know that EVALITA 2020 would not have been possible without the tireless effort, enthusiasm, and originality of the task organizers, the colleagues and friends involved in running them, and all the participants who contributed to make the workshop a success.

References

Chiara Alzetta, Alessio Miaschi, Felice Dell'Orletta, Frosina Koceva, and Ilaria Torre. 2020. PRELEARN@EVALITA2020: Overview of the prerequisite relation learning task for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. DIACR-Ita@EVALITA2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Pierpaolo Basile, Marco Lovetere, Johanna Monti, Antonioa Pascucci, Federico Sangati, and Lucia Siciliani. 2020b. Ghigliottin-AI@EVALITA2020: Evaluating artificial players
for the language game "la ghigliottina". In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Cristina Bosco, Silvia Ballarè, Massimo Cerruti, Eugenio Goria, and Caterina Mauri. 2020. KIPoS@EVALITA2020:overview of the task on kiparla part of speech tagging. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

[7] https://www.amazon.science/
[8] https://www.bnova.it
[9] https://www.celi.it/
[10] http://elra.info/en/
[11] https://research.google/

Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, and Roberto Zamparelli. 2020. AcCompl-it@EVALITA2020: Overview of the acceptability & complexity evaluation task for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

T. Caselli, N. Novielli, V. Patti, and P. Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)* , Turin, Italy. CEUR.org.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. SardiStance@EVALITA2020: Overview of the task on stance detection in italian tweets. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Andrea Cimino, Felice Dell'Orletta, and Malvina Nissim. 2020. TAG-it@EVALITA2020: Overview of the topic, age, and gender prediction task for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 2020a. CHANGE-IT@EVALITA2020: Change headlines, adapt news, generate. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Lorenzo De Mattei, Graziella De Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, and Giulia Rambelli. 2020b. ATE_ABSITA@EVALITA2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI@EVALITA2020: Automatic misogyny identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Lorenzo Gregori, Maria Montefinese, Daniele P. Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. CONCRETEXT@EVALITA2020: The concreteness in context task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Stefano Menini, Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2020. DaDoEval@EVALITA2020: Same-genre and cross-genre dating of historical documents. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Martina Miliani, Giulia Giorgi, Ilir Rama, Guido Anselmi, and Gianluca E. Lebani. 2020. DANKMEMES@EVALITA2020: The memeing of life: memes, multimodality and politics. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the evalita 2020 hate speech detection task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.