Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 174–180
Marseille, 20-25 June 2022

Italian NLP for Everyone: Resources and Models from EVALITA to the European Language Grid

Valerio Basile*, Cristina Bosco*, Michael Fell*, Viviana Patti*, Rossella Varvara†
*University of Turin, Italy
{valerio.basile,cristina.bosco,michaelkurt.fell,viviana.patti}@unito.it
†University of Fribourg, Switzerland
rossella.varvara@unifr.ch

Abstract

The European Language Grid enables researchers and practitioners to easily distribute and use NLP resources and models, such as corpora and classifiers. We describe in this paper how, during the course of our EVALITA4ELG project, we have integrated datasets and systems for the Italian language. We show how easy it is to use the integrated systems, and demonstrate in case studies how seamless the application of the platform is, providing Italian NLP for everyone.

Keywords: Italian, European Language Grid, EVALITA

1. The EVALITA4ELG Project

In the past 20 years the Natural Language Processing (NLP) area has been featured by a growing interest for standard formats for linguistic resources and for initiatives devoted to improve the accessibility and sharing of them. While the former have led to large and international projects, e.g., Universal Dependencies for morpho-syntactic annotation, the latter resulted in a wider availability and usability of resources on platforms such as META_SHARE[1], European Language Grid (ELG)[2], CLARIN[3], and DARIAH[4].

In this paper, we describe a recent initiative developed within the context of ELG, namely the EVALITA4ELG project, and we show how easy it is to use the systems and resources integrated on ELG for several NLP tasks. ELG, launched in early 2021 and funded by the European Union's Horizon 2020 research and innovation programme, is a large technological infrastructure aiming to become the primary platform for Language Technology in Europe. As a pilot funded by ELG itself, the EVALITA4ELG project aimed at systematically collect the resources and systems released for the EVALITA evaluation campaign, and make them easily accessible through the ELG platform.

EVALITA[5] is an evaluation campaign centered on a variety of phenomena and tasks for the Italian language. Started in 2007, it has been organized every two years ever since, with the contribution of the national and international community working on NLP for the Italian language. Similar campaigns are supported by other communities working on other languages, such as the *Workshop on Evaluation of Human Language Technologies for modern Iberian languages*[6] (IberEval), which started in 2010, but has only been held on a regular annual basis since 2017, and evolved into *Iberian Languages Evaluation Forum* (IberLEF); the *Natural Language Processing shared tasks for German*[7] (GermEval), which started in 2014; the *Défi Fouille de Textes*[8] (DEFT) for French, which started in 2005 as an annual event. Considering the diffusion of evaluation campaigns centered on other national languages, the expected impact of EVALITA4ELG is not only on the community focused on Italian, but also on other communities, for whom the project can serve as a model.

The growing number of tasks and related resources and services developed and shared during the EVALITA campaigns has made the need for a platform especially urgent, to organize and share the huge amount of data and tools with a larger set of scholars. The number of EVALITA tasks has grown from 5 in the first edition in 2007, to 10 in 2018, up to 14 tasks in the last edition, held in 2020 (Basile et al., 2020). The nature of the tasks also evolved, including a growing variety of exercises oriented to semantics and pragmatics, but without neglecting the more classical ones, like PoS tagging and parsing applied on an increasing variety of text genres and domains. In this, EVALITA follows the trends of other national and international evaluation campaigns, in particular SemEval[9]. On the other hand, an increasing amount of participants (51 teams from 14 countries in the 2020 edition) shows the peculiar vitality of the research community behind this campaign, which involves scholars from academy and industry, from Italy and foreign countries alike.

[1] http://www.meta-share.org/
[2] https://www.european-language-grid.eu/
[3] https://www.clarin.eu/
[4] https://www.dariah.eu/
[5] https://www.evalita.it/
[6] https://sites.google.com/view/ibereval-2018
[7] https://germeval.github.io/
[8] https://deft.limsi.fr/index.php?id=1
[9] https://semeval.github.io

2. Italian NLP Resources and Corpora

In the last year, the number of resources for the Italian language available on the ELG platform has grown consistently: from 13 resources to more than 300, including corpora and lexical resources. The project EVALITA4ELG contributed to this growth with the integration of 44 resources that were originally created and distributed in the context of the EVALITA evaluation campaign.

The project goal of integrating Italian linguistic resources into the ELG platform not only contributes to widening the representation of the Italian language in the platform, but it also highlights the contribution of the EVALITA community to the development of NLP resources, offering a space where data can be stored, documented and easily retrieved for future use and for reproducibility. The project aimed at preserving the efforts done by the organizers in creating benchmarks and resources for the evaluation of the tasks. In fact, existing resources frequently do not suit well for specific tasks, or they need additional layers of annotation. In this perspective, the integration of the resources into ELG provides multiple added values: i) it offers a unique place to collect together the resources developed in the context of EVALITA; ii) it promotes the EVALITA resources to users that were not aware of them before; iii) it directly provides detailed documentation of the releases of the resources.

2.1. Pre-existing Repositories

Traditionally, the release of benchmarks and data used for EVALITA tasks is curated by the organizers of the individual tasks; today, some of these repositories are no longer maintained, and frequently the links do not work anymore, hindering the data recovery and the subsequent replication of studies. In recent editions, a Github repository was created to keep all the data for a specific edition in a single place, but detailed information about the corpora and benchmarks should be retrieved from separate sources. Further, the EVALITA4ELG integration will encourage the reuse of resources for new tasks and goals.

2.2. Treatment of Sensitive Data

Some resources were not uploaded due to the particular nature of the data used, mainly for copyright issues. 13 resources required an anonymization procedure to mask personal and sensitive data before their distribution. Together with our partner CELI[10], we made use of an automatic anonymization tool developed for the AnonymAI H2020 research project ("NGI Trust", grant agreement n.825618), with a human-in-the-loop approach. We masked personal identifiers, including person names, emails, URL and Twitter mentions of non-public figures, to protect their privacy and comply with current regulations, while aiming at keeping the resources as valid as possible for benchmarking purposes.

[10] https://innovation.h-farm.com/en/celi/

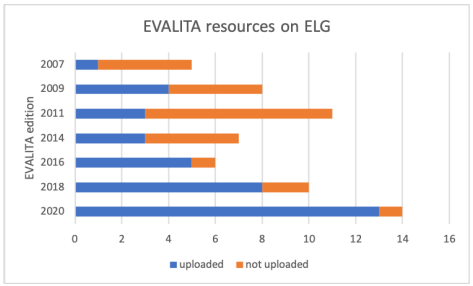

Figure 1: Number of EVALITA resources integrated in ELG per edition.

2.3. Collection Statistics

For each resource, we added a short description, keywords and its intended applications. The names and affiliations of the resource creators were listed as well. All the information were taken from the respective papers in the EVALITA proceedings, that were cited as well in the documentation field. Whenever available, a link to the official webpage of the task was added. The size of the dataset, its format and license of use were also described. The ELG infrastructure offers the possibility to describe the resource more in depth, e.g., adding information on the annotation procedure and the agreement. However, it requires only few obligatory fields, keeping the resource integration quick and easy. We uploaded 43 corpora and 1 lexical/conceptual resource (the CONcreTEXT dataset from EVALITA 2020 (Gregori et al., 2020)). For the latest editions (2016, 2018, and 2020), we were able to integrate around the 80% of the datasets; for previous editions it was more difficult to retrieve data, as can be seen in Figure 1 which shows the amount of resources broken down by year.

The uploaded corpora represent different language genres. In the first editions, tasks have mainly used balanced or general reference corpora, as well as news corpora. Since 2009, data from computer-mediated communications started being used, with Wikipedia articles (Attardi and Simi, 2009; Bos et al., 2009). Later on, in 2014, social media data started being used (Basile et al., 2014; Basile et al., 2018), and their use increased during the latest editions. Other less frequently employed types of linguistic data are transcribed spoken language data (NERTBN 2011 (Bartalesi Lenzi et al., 2013), iLISTEN 2018 (Basile and Novielli, 2018), KIPoS 2020 (Bosco et al., 2020)) and non-naturalistic data, i.e., made up texts like those used in the AcCompl-it task (Brunato et al., 2020). The varieties of textual genres just mentioned has been exploited in some tasks to test the adaptation of the systems to different domains. Frequently indeed, the performance of a system in a domain different from the trained one decreases. Thus, some data sets represent

multiple genres, in order to check for the flexibility of the systems. In the majority of EVALITA tasks, the focus has been on textual data of various genres, but a consistent group (13 out of 62 tasks) made use of speech data, and one task, DankMemes (Miliani et al., 2020), provided a multimodal dataset of images and text data, specifically, Internet memes. Table 1 shows the most popular intended applications of the corpora uploaded to the ELG, according to the metadata provided with the resources. Note that a single corpus may be associated with multiple intended application tags.

Frequency	Tag
14	*SentimentAnalysis*
5	*Parsing*
4	memes recognition
4	memes classification
4	*PolarityDetection*
4	*ImageUnderstanding*
4	*DependencyParsing*
3	*IronyDetection*
3	*ConstituencyParsing*
68	other

Table 1: Distribution of the "Intended Applications" tag of corpora. Highlighted tags are from the ELG metadata ontology, while the others are custom tags.

3. Italian NLP Models and Systems

We added new service functions to the ELG portfolio, based on the models built by EVALITA participants. The models are of different kinds. In this section, we describe the models that we have integrated into the ELG (see Section 3.1) and the deployment process to integrate new models into the ELG (see Section 3.2)).

3.1. Models integrated in the ELG

We integrate systems as a pool of web services with a common interface, deployed on a dedicated hardware infrastructure. The APIs are accessible online with terms and conditions compatible with the ELG best practices. It is also possible to try out the services through the ELG platform[11]. Resources and services integrated in the ELG in the context of EVALITA4ELG are accessible through the ELG webpage[12]. In our proof of concept, we have integrated 8 systems into the ELG, including pre-neural models with tf-idf vector-based feature representation. Others are neural networks, or even general transformer language models (e.g., AlBERTo for Italian social media language[13]).

We have made available via the ELG platform systems for the following tasks: KiPOS (part-of-speech tagging, 1 system), AMI (misogyny identification, 1 system), HaSpeeDe-2018 (hate speech detection, 3 systems), HaSpeeDe-2020 (hate speech detection, 2 systems), GxG (gender prediction, 1 system). The use of the systems is easy. Holders of ELG accounts can query those services by simple POST calls to the relevant API, where a text is being sent as input, and the API response contains the classification of the text (for classifiers), or the tagged sequence (for sequence tagging tasks).

3.2. Deploying systems

For each system, we defined and followed a specific procedure to integrate them into the ELG. All systems are recent, from 2018 or 2020, and we found an overwhelming bias towards using Python as programming language in the system implementations. This made the integration process more uniform, as besides small differences, we used the same pattern to integrate all systems into the ELG platform. Specifically, we used the following as typical steps to integrate an EVALITA system into the ELG:

1. Obtain the system source code and the training data. Usually, it is necessary to train the system, while rarely a pre-trained system can be obtained.
2. Make the code runnable. As the code is for an EVALITA task, there is typically code to train, test, and evaluate. A model can be trained and evaluated to check if it produces similar results to what was reported.
3. Train the system and save the trained model, as for the ELG integration, we only need the inference stage. Remove anything from the code that is not test, specifically training and evaluation code.
4. Adapt the interface of the code to take as input the conforming ELG JSON format and output the conforming JSON format as well.[14]
5. Dockerize the system. For this purpose, the Dockerfile and entrypoint provided in the template can be used as a starting point.
6. To make the specific implementation run, specify a requirements.txt file where all Python packages necessary are defined. The command `pip freeze > requirements.txt` can be used in a working Python environment to find out the required packages. The more specific the package version numbers are, the better the chances that the docker image will stand the test of time.

[11] See for example the Italian POS Tagger for spoken language (SoMeWeTa), developed in the context of the KIPoS task at EVALITA 2020 (Proisl and Lapesa, 2020): https://live.european-language-grid.eu/catalogue/#/resource/service/tool/5214.

[12] https://live.european-language-grid.eu/catalogue/#/resource/projects/1397

[13] https://github.com/marcopoli/AlBERTo-it

[14] In the case of Python code, this can be achieved by adapting an available template: https://gitlab.com/european-language-grid/usfd/simple-python-tokeniser

7. Specify the meta data of the system on the ELG website (this is a very similar to the one for resources/corpora described in the previous section). An important parameter to specify here is the location of the docker image.

8. Host the docker image in a controlled environment for testing (e.g., Dockerhub or Gitlab). If the platform allows it, continuous integration is a useful tool to automatically re-build the docker image whenever the code is updated.

9. After the system is submitted, including all the metadata, the ELG staff validates the entry and reaches out in case there is something not working or some information missing.

Frequency	Tag
12	*TextCategorization*
11	*HateSpeechRecognition*
1	*PosTagging*
68	other

Table 2: Distribution of the "Intended Applications" tags across the tools.

Table 2 shows the intended applications of the tools integrated into the ELG. Note that a single tool may be associated with multiple intended application tags, and that the table includes multiple versions of the tools.

4. Validation by Case Studies

One of the main advantages of implementing EVALITA-based models and datasets as services in the ELG is the availability of a uniform, easy to use interface to access them, directly but also programmatically. In order to validate this claim, and therefore the work described in this paper, we propose two experiments on EVALITA benchmarks where the added value is the ability to easily combine datasets and models and build more sophisticated systems out of the rich collection of resources made available by the EVALITA4ELG project. In the rest of this section, we show that just a few lines of code are sufficient to create systems for recent NLP tasks on the Italian language, thanks to the ELG infrastructure.

4.1. Experiment 1: Multi-task Learning

In the first experiment, our goal is to check the feasibility of creating new systems that leverage the variety of data imported into the ELG by ELG4EVALITA. We selected two corpora, both annotated at the message level, and performed a multi-task learning (MTL) classification experiment. The corpora are HaSpeeDe 2 (Sanguinetti et al., 2020) from EVALITA 2020 (ELG ID: 7498), and in particular the task A (Hate Speech Detection), and IronITA (Cignarella et al., 2018) from EVALITA 2018 (ELG ID: 7372), and in particular task A (Irony Detection). Both corpora are annotated with binary labels indicating the presence or absence of hate speech and irony in the natural language, respectively. For both tasks, a training set and a test set are provided. Downloading the data is an extremely simple process, achieved through the `elg` Python package available from the public repository[15]:

```
pip install elg
```

The data can be downloaded directly from a Python script, by specifying the resource ID. For instance, for the HaSpeeDe 2 corpus, the following two lines of code download a ZIP archive containing all the data of the task:

```
data_hs = Corpus.from_id(7498)
data_hs.download()
```

Following the data harmonization effort carried out during the project, the two datasets are presented in a comma-separated format (CSV) with the same structure and field names (except for the labels, which encode different phenomena). It is therefore straightforward to read both annotated corpora in one go and encode them with the same procedure. In particular, for the texts we used the *Wordpiece-based* tokenizer provided by AlBERTo, the BERT model pre-trained on Italian social media data (Polignano et al., 2019).
In order to build a MTL model, we concatenated the two training sets setting the labels to -1 when absent in the original data, i.e., the irony label for the HaSpeeDe 2 data and the hate speech label for the IronITA data. We then fine-tuned AlBERTo on the joint dataset with a custom loss function implementing Masked Binary Crossentropy, which is a straightfoward variation on Binary Crossentropy where the labels having value -1 are ignored in the computation of the loss value. In other words, we created a *hard parameter sharing* MTL model and trained it to learn the hate speech and irony labels jointly.
Table 3 presents the results of the experiment on the HaSpeeDe 2 test set, that is, classification of Hate Speech. The figures are the average of five runs, and show how the additional information coming from IronITA boosts the performance of the classifier on the HaSpeeDe 2 benchmark, comparing to fine-tuning the vanilla AlBERTo model. We used a batch size of 32, a fixed learning rate of $10^{-}5$ and 5 training epochs for both the baseline model and the MTL model.
While reaching a top classification performance was outside the scope of this paper, it should be noted that the overhead in terms of coding was minimal, and the development process could be focused solely on creating the MTL model rather than searching, accessing, and harmonizing the data.

4.2. Experiment 2: Ensemble Systems

Within the technological framework provided by ELG, creating ensemble models is rather straightforward. We

[15] https://pypi.org/project/elg/

Model	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	P_{macro}	R_{macro}	$F1_{macro}$	Acc.
AlBERTo	.849	.666	.738	.724	.867	.784	.786	.767	.761	.765
$\text{MTL}_{\text{irony}}$	.772	.795	.779	.786	.749	.762	.779	.772	**.771**	**.772**

Table 3: Multi-task learning results on HaSpeeDe 2.

```
hansel = Service.from_id(7370)
montanti = Service.from_id(5315)
models = [hansel, montanti]
text = "this is an example sentence"
results = []
for model in models:
    results.append(model'(request_input=text, request_type="text"))
    % model' returns 1 or 0, depending on classification result
ensemble_label = 1 if sum(results) == 2 else 0
% example for two models, ensemble label is 1 if both labels are 1
return ensemble_label
```

Figure 2: Python code for the ensemble model.

Model	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	P_{macro}	R_{macro}	$F1_{macro}$	Acc.
hs-rug	.774	.694	.732	.774	.694	.732	.745	.743	.742	.742
hs-fbk1	.769	.697	.732	.716	.785	.748	.742	.741	.740	.740
hs-hansel	.768	.715	.741	.726	.778	.751	.747	.746	.746	.746
hs-montanti	.741	.761	.751	.747	.725	.736	.744	.743	.743	.743
$\text{ensemble}_{\geq 1}$	.870	.510	.643	.646	.921	.759	.758	.716	.701	.713
$\text{ensemble}_{\geq 2}$	.810	.660	.727	.706	.841	.767	.758	.750	.747	.749
$\text{ensemble}_{\geq 3}$	.752	.799	.775	.778	.728	.752	.765	.764	**.764**	**.764**
$\text{ensemble}_{=4}$	.692	.899	.782	.849	.588	.695	.771	.744	.739	.746

Table 4: Result of the individual hate spech models and different configurations of the ensemble model. The performance is measured in terms of Precision (P), Recall(R), and F1-score (F1) for the negative class (0), the positive class (1), and their arithmetic mean (macro).

built a simple parametric ensemble starting from four different hate speech detection systems presented at the HaSpeeDe tasks. Specifically, we retrieve predictions for each single system, then we aggregate the labels using a simple voting mechanism. If more than N systems label a message as containing hate speech, then that is the final label assigned to the target message. For instance, with $N = 3$, the ensemble with strictly more than two single votes, $\text{ensemble}_{\geq 3}$ represents the majority vote. Figure 2 shows the Python code required to retrieve the single system responses and combine them into a simple ensemble voting model. To allow a simpler presentation, the code shows the example of $\text{ensemble}_{\geq 2}$ for only two systems. The core logic is completely encapsulated into the function call of $model$, and the harmonization of the input and output formats allows for a very compact code. Note that $model'$ parses the JSON response of $model$ to obtain the final 1 or 0 as response.

We tested the models, which are trained on the HaSpeeDe 2 training data, on the HaSpeeDe 2 test set, specifically on task A (hate speech detection) on Twitter data. Table 4 shows the results of this experiment, with the four different systems, four values for the voting, highlighting how even a very simple ensembling technique is capable of improving the performance of individual models (in this case, with $N = 3$).

# of resources	Resource type
42	Corpus
1	LexicalConceptualResource
13	ToolService

Table 5: Summary of the resources uploaded into the ELG by the EVALITA4ELG project, by type, including multiple versions.

5. Conclusion

The EVALITA4ELG project brought a great deal of language resources for Italian into the European Language Grid, improving their accessibility and discoverability. Table 5 summarizes the results of this effort. Besides annotated corpora, the platform now offers easy access to trained models for a variety of tasks, and an interface to combine them and build more refined systems. The know-how and the procedures developed during the project will pave the way for similar initiatives for other languages.

Acknowledgements

This work was funded by ELG (European Language Grid) Pilot Projects Open Call 1 (Grant Agreement No. 825627 – H2020, ICT 2018-2020 FSTP). Project website: http://evalita4elg.di.unito.it/

6. Bibliographical References

Attardi, G. and Simi, M. (2009). Overview of the EVALITA 2009 Part-of-Speech tagging task. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, volume 9, Reggio Emilia, Italy, December.

Bartalesi Lenzi, V., Speranza, M., and Sprugnoli, R. (2013). Named Entity Recognition on Transcribed Broadcast News at EVALITA 2011. In Bernardo Magnini, et al., editors, *Evaluation of Natural Language and Speech Tools for Italian. International Workshop, EVALITA 2011, Rome, January 24-25, 2012, Revised Selected Papers*, pages 86–97. Springer.

Basile, P. and Novielli, N. (2018). Overview of the evalita 2018 italian speech act labeling (ilisten) task. In Tommaso Caselli, et al., editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263, Torino, December. CEUR Workshop Proceedings (CEUR-WS. org).

Basile, V., Bolioli, A., Patti, V., Rosso, P., and Nissim, M. (2014). Overview of the EVALITA 2014 SENTIment POLarity Classification task. In Cristina Bosco, et al., editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 50–57. Pisa University Press, December.

Basile, V., Novielli, N., Croce, D., Barbieri, F., Nissim, M., and Patti, V. (2018). Sentiment Polarity Classification at EVALITA: Lessons Learned and Open Challenges. *IEEE Transactions on Affective Computing*, 12(2):466–478.

Basile, V., Croce, D., Maro, M. D., and Passaro, L. C. (2020). EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765, online. CEUR Workshop Proceedings (CEUR-WS. org).

Bos, J., Zanzotto, F. M., and Pennacchiotti, M. (2009). Textual entailment at evalita 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, volume 9, Reggio Emilia, Italy, December.

Bosco, C., Ballarè, S., Cerruti, M., Goria, E., and Mauri, C. (2020). KIPoS@EVALITA2020: Overview of the Task on KIParla Part of Speech tagging. In Valerio Basile, et al., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR Workshop Proceedings (CEUR-WS. org).

Brunato, D., Chesi, C., Dell'Orletta, F., Montemagni, S., Venturi, G., and Zamparelli, R. (2020). AcCompl-it@EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian. In Valerio Basile, et al., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR Workshop Proceedings (CEUR-WS. org).

Gregori, L., Montefinese, M., Radicioni, D. P., Ravelli, A. A., and Varvara, R. (2020). CONCRETEXT@EVALITA2020: The concreteness in context task. In Valerio Basile, et al., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR Workshop Proceedings (CEUR-WS. org).

Miliani, M., Giorgi, G., Rama, I., Anselmi, G., and Lebani, G. E. (2020). DANKMEMES@EVALITA2020: The memeing of life: memes, multimodality and politics. In Valerio Basile, et al., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR Workshop Proceedings (CEUR-WS. org).

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Proisl, T. and Lapesa, G. (2020). KLUMSy@ KIPoS: Experiments on Part-of-Speech Tagging of Spoken Italian. In Valerio Basile, et al., editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR Workshop Proceedings (CEUR-WS. org).

7. Language Resource References

Cignarella, A. T., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P., et al. (2018). Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.

Sanguinetti, M., Comandini, G., di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V.,

and Russo, I. (2020). Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task. In Valerio Basile, et al., editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765. CEUR Workshop Proceedings (CEUR-WS.org). Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2020 ; Conference date: 17-12-2020.