EVALITA4ELG: Italian Benchmark Linguistic Resources, NLP Services and Tools for the ELG Platform

Viviana Patti, Valerio Basile,
Cristina Bosco, Rossella Varvara,
Michael Fell*
Università degli Studi di Torino

Andrea Bolioli, Alessio Bosca**
CELI - Language Technology

Starting from the first edition held in 2007, EVALITA is the initiative for the evaluation of Natural Language Processing tools for Italian. This paper describes the EVALITA4ELG project, whose main aim is at systematically collecting the resources released as benchmarks for this evaluation campaign, and making them easily accessible through the European Language Grid platform. The collection is moreover integrated with systems and baselines as a pool of web services with a common interface, deployed on a dedicated hardware infrastructure.

1. Introduction

Scientific investigations are universally characterized by the ability to validate models and research outcomes through the rigor of quantitative measures. This requires, on the one hand, a precise definition of the research problem and focus and, on the other hand, the definition of an evaluation framework to compare different ideas and approaches to solutions. In the field of Natural Language Processing (NLP), periodical benchmarking campaigns challenge systems on problems inherent to semantics (e.g., machine translation and question answering) or syntax (e.g., parsing). Undoubtedly, these campaigns are drivers of technological progress and interest for classic and novel problems rooted on specific language processing phenomena. Shared tasks are, nowadays, increasingly popular, as a common tool in the NLP community to set benchmarks for specific tasks and promote the development of comparable systems. They "revolve around two aspects: research advancement and competition", being the research advancement the driving force and main goal behind organizing them (Nissim et al. 2017). Moreover, just as importantly, they often help provide a boost to under-researched and under-resourced topics and languages.

The application of existing NLP methods to different data sets and languages is a crucial aspect to track advances in the field and to assess the impact of the work done in the community. The validation of existing NLP models, indeed, strongly depends on the possibility of generalizing their results on data and languages other than those on which they have been trained and tested (Magnini et al. 2008). Moving forward along this line, most recent trends are pushing toward proposing benchmarks for multiple tasks

* Dept. of Computer Science - C.so Svizzera 185, 10149, Turin, Italy.
E-mail: name.surname@unito.it

** Via San Quintino, 31 10121, Turin, Italy
E-mail: name.surname@celi.it

(Wang et al. 2018), or for testing adaptability of systems to different textual domains and genres, or to different languages. The recent specific emphasis on multilingual assessment is also driven by a growing community awareness that language technologies can help promote multilingualism and linguistic diversity (Joshi et al. 2020). However, much of the research still focuses on a few dominant languages such as English, despite the fact that there is easy access to user-generated content in many languages. In this perspective lies the EVALITA4ELG project for the integration of linguistic resources and language technologies developed in the context of the EVALITA evaluation campaign into the European Language Grid[1]. The ELG is a new growing platform for Language Technology in Europe funded by Horizon 2020 (Rehm et al. 2020a), with the ultimate goal of creating an open and shared linguistic benchmark for Italian on a large set of representative tasks.

EVALITA is an initiative of the Italian Association for Computational Linguistics (Associazione Italiana di Linguistica Computazionale, AILC[2]). Since 2007, it provides a shared framework where different systems and approaches can be evaluated and compared with each other with respect to a large variety of tasks, organized by the Italian research community. The proposed tasks represent scientific challenges where methods, resources, and systems are tested against shared benchmarks representing linguistic open issues or real world applications. The focus of EVALITA is to support the advancement of methodologies and techniques for natural language and speech processing in an historical perspective, beyond the performance improvement — favoring reproducibility and cross-community engagement — and on exploring the multilingual and multi-modal dimensions.

The Italian language is underrepresented in the ELG platform, currently including few NLP services and corpora and related to a very limited number of NLP tasks. It mostly includes parallel corpora for machine translation and multilingual dependency treebanks, focused on texts featured by standard forms and syntax. However, several resources and models were developed in the past decades for the Italian language which also cover multiple applications, domains, text types and genres. Examples of the prolific activity carried out by the Italian NLP community are precisely the results provided in the context of the EVALITA campaigns. Our main aim is to build the catalogue of EVALITA resources and tasks ranging from traditional tasks like POS-tagging and parsing to recent and popular ones such as sentiment analysis and hate speech detection on social media (Basile et al. 2017), and integrate them into the ELG platform. The project includes the integration of state-of-the-art Language Technology (LT) services into the ELG platform, accessible as web services.

We aim at leveraging over a decade of output of the Italian NLP community, providing through the ELG platform an easier access to the resources and tools developed over the seven editions of EVALITA. This will constitute a precondition to develop a framework for evaluating the performance of models across a diverse set of tasks in order to fully exploit the ELG potential to be a community reference for resources, benchmarks, services, going beyond the mere catalogue.

In the following, we will briefly introduce, the ELG platform and the EVALITA evaluation campaign, in a historical perspective. Then, we will illustrate the EVALITA4ELG goals and the first outcomes, with a special focus on the new opportunities that are opening up for the community.

1 https://www.european-language-grid.eu/
2 https://www.ai-lc.it/

2. Background

In this Section, we describe the two projects on which our EVALITA4ELG project is fundamentally built on: the European Language Grid platform (ELG, see Section 2.1) and the Evaluation Campaign for NLP and Speech Tools for Italian (EVALITA, see Section 2.2).

2.1 The European Language Grid Platform

The European Language Grid (ELG) is a three-year H2020-EU project (2019-2022)[3] whose aim is to establish the primary platform and marketplace for the European Language Technology community, both industry and research (Rehm et al. 2020a).

The history of the ELG dates back to 2010 and to the original META-NET[4] cluster of projects (T4ME, CESAR, METANET4U, META-NORD). The first ideas and concept of a "European Service Platform for Language Technologies" were published in the META-NET Strategic Research Agenda for Multilingual Europe 2020 (Rehm and Uszkoreit 2013). Since then, many projects and initiatives have been carried out in Europe in order to create multilingual repositories of linguistic resources, data and services (Rehm et al. 2020b). Over 400 linguistic resources (LRs) from ELRA[5], ELRC-SHARE[6] and META-SHARE[7] have been integrated into ELG Release 1 (April 2020). Concerning the language coverage, the platform includes resources for all EU languages, both EU official and EU candidate languages, e.g., Catalan, Czech, English, French, Frisian, German, Greek, Italian, Hungarian, Portuguese, Romanian, Russian, Serbian. Moreover, it includes also LRs for non-EU languages, such as languages spoken by EU immigrants or languages of political and trade partners, e. g., Amharic, Arabic, Chinese, Hindi, Mongolian, Nepali, Pashto, Persian.

The ELG is developed to be a scalable cloud platform providing access to hundreds of commercial and non-commercial LTs for all European languages, including running tools and services as well as data sets and resources. The platform has an interactive web user interface and corresponding backend components and REST APIs. It offers access to various kinds of resources such as corpora and data sets as well as functional LT services, i.e., existing LT tools that have been containerised and wrapped with the ELG LT Service API. The platform backend contains the ELG catalogue, i.e., the central list of metadata records of functional services, nonfunctional resources (e.g., data sets), but also the entries of organisations (e. g., companies, universities, research centres) and other stakeholders, as well as service types, languages and other types of information. All entities are described in compliance with the ELG-SHARE metadata schema[8] (Labropoulou et al. 2020). ELG is collaborating with other related projects and initiatives that share this approach, such as AI4EU, ELRC, CLAIRE, CLARIN, HumanE-AI, and others.

3 ELG - European Language Grid, Grant agreement ID: 825627 funded under H2020-EU.2.1.1., https://cordis.europa.eu/project/id/825627
4 META-NET: A Network of Excellence forging the Multilingual Europe Technology Alliance, http://www.meta-net.eu/
5 http://www.elra.info/en/catalogues/catalogue-language-resources/
6 https://elrc-share.eu/
7 http://www.meta-share.org/
8 ELG-SHARE metadata schema, https://gitlab.com/european-language-grid/platform/ELG-SHARE-schema

In April 2020 ELG published the first open call, which resulted in 110 project proposals submitted for evaluation and funding. A total of 10 pilot projects were selected and financially supported. EVALITA4ELG is one of these projects.

2.2 EVALITA: Evaluation of NLP and Speech Tools for Italian

Starting from the first edition held in 2007, EVALITA[9] has been proposed as the initiative devoted to the evaluation of Natural Language Processing tools for Italian, providing a shared framework where participating systems had the possibility to be evaluated on a growing set of different tasks. EVALITA is a biennal initiative of AILC (Associazione Italiana di Linguistica Computazionale), and since 2016 has been co-located with CLiC-it, the Italian Conference on Computational Linguistics.

Some other evaluation initiatives are mostly focused on particular task typologies, let us mention the CoNLL shared tasks[10] for parsing, the SemEval workshop series on Semantic evaluation[11] or TREC[12] for information retrieval, regardless of the involved language (and not necessarily in a multilingual perspective). On the contrary, EVALITA is an evaluation campaign, which rather than being focused on specific language processing phenomena, is centred on a variety of phenomena and tasks for a single language, namely Italian. This allowed to gather around the initiative the national and international community working on NLP for the Italian language. Even if, given the campaign focus, in most cases data sets and tasks concerned Italian only, in the last editions there were few proposed tasks which comprehended evaluation also for other languages, fostering cross-linguistic evaluations but always keeping Italian as core language. This is the case of *Automatic Misogyny Identification task* (AMI) held in EVALITA 2018 (Fersini, Nozza, and Rosso 2018), which provided data and evaluation for both Italian and English languages, and of other shared tasks, as emerged from the first outcomes of our survey (see Section 5.2 for details).

Establishing shared standards, resources, tasks and evaluation practices with reference to languages other than English is a fundamental step towards the continued development of NLP (Magnini et al. 2008). In line with EVALITA, other events have been and are being organized by other national communities: the *Workshop on Evaluation of Human Language Technologies for modern Iberian languages*[13] (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)[14] since 2019[15]; the *Natural Language Processing shared tasks for German*[16] (GermEval), which started in 2014; the *Défi Fouille de Textes*[17] (DEFT), which started in 2005 as an annual event. Nevertheless, most of such campaigns in each edition have focused on a single task or a small number of tasks, while the EVALITA's perspective has always been characterized by its openness to the

9 http://www.evalita.it/
10 The SIGNLL Conference on Computational Natural Language Learning, https://www.conll.org/
11 The International Workshop on Semantic Evaluation https://semeval.github.io/
12 The Text REtrieval Conference, https://trec.nist.gov/
13 IberEval 2018: https://sites.google.com/view/ibereval-2018, http://nlp.uned.es/IberEval-2017/index.php, http://gplsi.dlsi.ua.es/congresos/ibereval10/
14 IberLEF 2021: https://sites.google.com/view/iberlef2021/)
15 Starting in 2019, the TASS and IberEval evaluation activities have joined forces and communities to set up the new evaluation forum IberLEF, maintaining the focus on Iberian languages, such as Spanish, Portuguese, Catalan, Basque and Galician.
16 https://germeval.github.io/
17 https://deft.limsi.fr/index.php?id=1

wider possible variety of tasks. This is also confirmed by the inclusion of evaluation exercises about speech: each edition of the EVALITA campaign, held in 2007 (Magnini et al. 2008), 2009 (AA.VV. 2009), 2011 (Magnini et al. 2013), 2014 (Attardi et al. 2015), 2016 (Basile et al. 2017), 2018 (Caselli et al. 2018), 2020 (Basile et al. 2020), has been organized around a set of shared tasks dealing with both written (most of the tasks) and spoken language, varying with respect to the challenges tackled and the data sets used.

Such variety has permitted to give to the interested scholars, in each edition of the event, a clearer assessment of both the distribution of NLP research groups in Italy and for Italian, and of the complexity of proposed tasks also with reference to the state of development of Italian linguistic resources (Magnini et al. 2008).
The number of tasks has considerably grown, from 5 tasks, in the first edition in 2007, to 10 tasks in 2018, and 14 tasks in the edition held in 2020 (Basile et al. 2020), showing the peculiar vitality of the research community behind this campaign, which involves scholars from academy and industry, from Italy and foreign countries alike. Following the trends of other national and international evaluation campaigns, like e.g. SemEval[18], the typology of tasks also evolved, progressively including a larger variety of exercises oriented to semantics and pragmatics, but without neglecting the more classical ones, like PoS tagging and parsing. In particular, edition 2016 brought an innovative aspect in the foreground, i.e. the focus on social media data, especially Twitter, and the use of shared data across tasks, yielding a test set with layers of annotation concerning PoS tags, sentiment information, named entities and linking, and factuality information (Basile et al. 2017). Organisers were more and more encouraged to collaborate, stimulated to the creation of a shared test set across tasks, and to eventually share all resources with a wider audience. This has resulted in the creation of GitHub public repositories where EVALITA resources can be accessible[19]. The 2016 edition saw also a greater involvement of industrial companies in the organisation of tasks. This can be seen as a reflection of the implementation of strategies, following the outcome of the survey and of the fruitful discussion that took place during the panel "Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign" held at CLiC-it 2015 (Sprugnoli, Patti, and Cutugno 2016). Testing adaptability of systems to different textual domains and genres also emerged as an important trend in recent campaigns, as evidenced by the proposal of shared tasks such as *Cross-Genre Gender Prediction* (GxG) in Italian (Dell'Orletta and Nissim 2018).

Another axis of evolution that characterizes the latest campaigns concerns the fruitful relationships that exist between EVALITA and other different evaluation campaigns that focus on different languages. For some very popular tasks, such as sentiment polarity classification, aspect-based sentiment analysis, stance detection, irony detection, hate speech detection, author profiling, entity linking, twin tasks have been organised in different campaigns like EVALITA, DEFT, SemEval and IberEval/IberLEF, CLEF possibly with extensions and variations. Let us mention, in this context, the HaSpeeDee task, proposed at EVALITA in 2018 (Bosco et al. 2018; Sanguinetti et al. 2020) and 2020 on different text genres (Twitter, FaceBook, newspaper headlines) for Italian, and reproposed with the HatEval shared task at SemEval 2019 (Basile et al. 2019) for English and Spanish on Twitter data, with a specific focus on cross-comparison between manifestations of hatred in different languages and with different hate targets, such as immigrants and women. This trend seems to create good conditions for the creation of a European

18 https://semeval.github.io
19 https://github.com/evalita2016/data

assessment framework, where benchmarking activities on different languages interact in a coordinated way, as also advocated by the promoters of the development of the ELG platform.

The comparison of the performance of systems against shared data sets and benchmarks improves the significance of the results obtained by participants and establishes acknowledged baselines, while the development of the data sets needed for training and/or testing is in itself an activity which positively influences the area, widening the availability of reliable resources to be exploited for both evaluation and application purposes (Attardi et al. 2015).

From its birth and until now, EVALITA is organized on a fully voluntary basis and allowed researchers to share and use resources on which state of the art for Italian NLP is defined. Since their systematic collection and sharing are among the goals of EVALITA from the beginning, by providing a platform for making resources and models more accessible the EVALITA4ELG project represents a meaningful improvement of the evaluation campaign. Open access to resources and research artifacts, such as data, tools, and dictionaries, is deemed crucial for the advancement of the state of the art in scientific research (Caselli et al. 2018) and the availability of shared evaluation benchmarks is crucial for fostering reproducibility and comparability of results.

3. EVALITA4ELG: Project Description and Goals

With the EVALITA4ELG project, we aim at leveraging over a decade of findings of the Italian NLP community, providing through the ELG platform an easier access to the resources and tools developed in the context of the EVALITA periodical benchmarking campaign for the Italian language. In this context, we work towards the achievement of multiples goals, namely: (i) a survey of the past 62 tasks organized in the seven editions of EVALITA, released as a knowledge graph; (ii) a common anonymization procedure of the resource data for improving their compliance with current standard policies; (iii) the integration of resources and systems developed during EVALITA into the ELG platform; (iv) the creation of a unified benchmark for evaluating Italian Natural Language Understanding (NLU) systems; (v) the dissemination of a shared protocol and a set of best practices to describe also new resources and new tasks in a format that allows a quick integration of metadata into the ELG platform.

As a first step, we started surveying the 62 tasks organized in the context of the EVALITA campaign. In this phase we aimed both at collecting the resources and their metadata for the following upload on the ELG platform, and at organizing this information into an ontology that we make available for it to be easily queried by the community. For this purpose, we structured a knowledge graph in the RDF-OWL format using Protégé, which is described in more details in Section 5. It will be available on the ELG platform for researchers, future organizers of tasks or stakeholders to, e.g., quickly search for a particular type of resource or find information about past evaluation campaigns. It can be used to extract trends among different editions of EVALITA in terms of tasks features, such as their application, language types, number of people and institution involved in the events, etc. In section 5.1, we show some examples of SPARQL queries and in section 5.2 we highlight some trends with regard to the resources employed and created during the EVALITA editions.

Second, we carefully checked whether the collected resources contain sensitive data that must be anonymized according to the current policies for the protection of people's privacy. The procedure for checking the resources and the methodology for

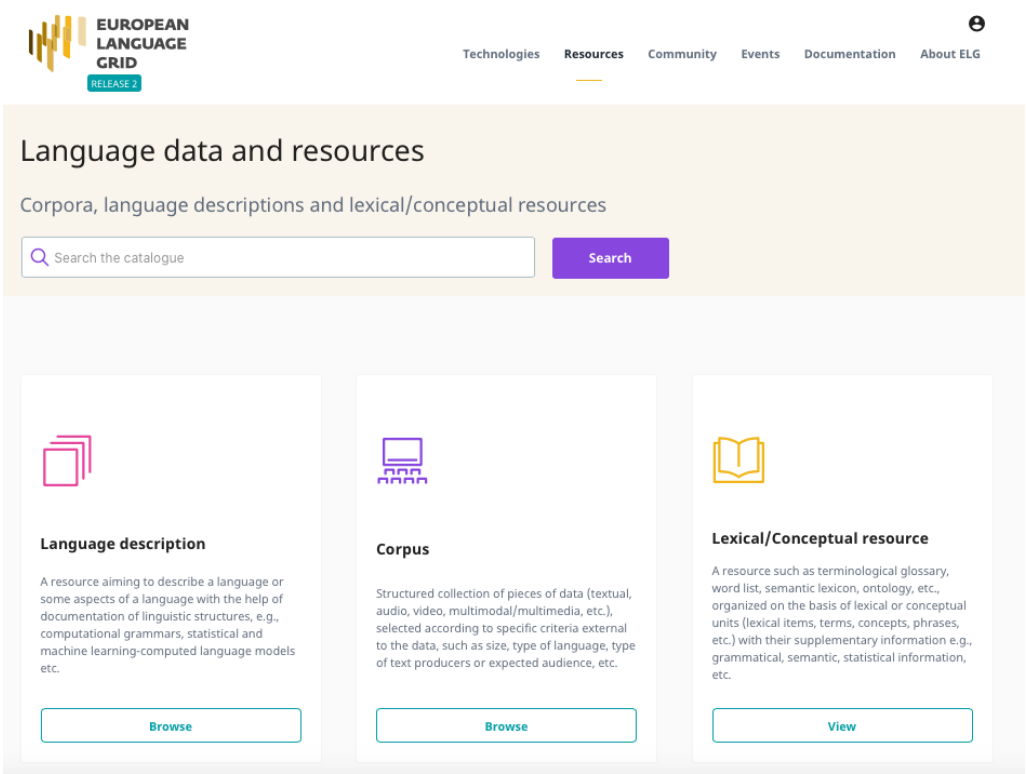

Figure 1
The ELG catalogue query website. It allows browsing and searching for different kinds of language resources: language descriptions, corpora, lexical/conceptual resources

anonymizing data were carried out by CELI[20], a company which is partner of University of Turin in the EVALITA4ELG project. We provide more information in Section 4 about this part of the project.

The integration of resources and systems developed during EVALITA into the ELG platform will allow a meaningful broadening of the ELG portfolio with corpora and resources for Italian, widening the ELG coverage with a larger spectrum of linguistic phenomena, language genres and domains (e.g., social media data). Once the resources are uploaded, users can access them through the online catalogue[21]. Items are classified according to their typology, as being a corpus, a lexical/conceptual resource (e.g., a semantic lexicon or an ontology) or a language description (e.g., a computational grammar or language model). Users can search for resources in the ELG catalogue by language and license filters. Figure 1 shows the ELG catalogue query website.

By selecting a resource, a set of metadata is shown, specifically a short description of the resource, its intended application, related keywords, and its documentation. In some cases, the data can be directly downloaded from the platform, otherwise a link to a website where the resource can be accessed or downloaded is provided. A first example

20 https://www.celi.it/en/
21 https://live.european-language-grid.eu/page/resources

of dataset uploaded from the EVALITA 2020 campaign is the KIPoS dataset[22] (Bosco et al. 2020). With the addition of the EVALITA resources, the ELG platform will thus become a trusted reference point, an open access for the research community to NLP resources for the Italian language, easing future research and significantly contributing to the dissemination of knowledge and outputs.

Moreover, we are adding into the ELG portfolio new service functions based on the best models built from EVALITA task participants and general transformer models (e.g., AlBERTo for Italian social media language[23]). We integrate systems and baselines as a pool of web services with a common interface, deployed on a dedicated hardware infrastructure. The APIs will be accessible online with terms and conditions compatible with the ELG best practices. It is also possible to try out the service through the ELG platform[24]. Resources and services integrated in ELG in the context of EVALITA4ELG project will be accessible through the ELG project webpage[25].

As a fourth point, we aim at laying the foundations for a multi-task benchmark platform for NLU in Italian, that can be also used in future EVALITA editions, on the line of what has been done for English with GLUE[26] (Wang et al. 2018). Given the variety of tasks and data sets in EVALITA, it will be possible to set up tools to evaluate models performance across multiple NLU tasks. NLU models, indeed, are frequently designed for a specific task, and their performance drops when tested on out-of-domain data. A multi-task benchmark will encourage the development of general and flexible NLU models. We will set up an evaluation platform and a leaderboard, favoring the focus of an overall picture of the state-of-the-art of language technologies for Italian.

Lastly, as a result of the EVALITA4ELG project, we will provide directions for the description of tasks and resources for the organizers of future editions of EVALITA. This will enable an easier integration of future EVALITA resources in the ELG platform. Drawing inspiration from the LREC online form proposed to authors of resource papers to standardise the descriptions of resources, we aim to provide the Italian community with a form to describe their resources and tasks. This goes in the direction to encourage the constant injection of the resources developed by the community into ELG in the future, supported from the effort of sharing best practices to be followed in the creation, description and release of resources for the easiest integration of data and metadata into the ELG platform. . We will encourage its use in future editions of EVALITA and in the co-located event CliC-it[27], especially in connection with the Resource and Evalution track.

4. Anonymization of Resources

All the EVALITA's resources to be made accessible in the ELG platform are carefully checked and made compliant with the current policies about data releasing and sharing (e.g., G.D.P.R. (Rangel and Rosso 2018)), even if originally created some years ago and

22 https://live.european-language-grid.eu/catalogue/#/resource/service/corpus/4894

23 https://github.com/marcopoli/AlBERTo-it

24 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.

25 https://live.european-language-grid.eu/catalogue/\#/resource/projects/1397

26 https://gluebenchmark.com

27 The Italian Conference on Computational Linguistics is the annual event organized by the Italian Association for Computational Linguistics (AILC, https://www.ai-lc.it/en/

sometimes following policies no longer valid. Particular attention has been paid to the anonymization of data.

The data sets collected for EVALITA4ELG were anonymized relying on an automatic anonymization tool developed in the context of *AnonymAI* research project, and then manually reviewed in order to assess their quality. The virtuous synergy with *AnonymAI* was important to accomplish this step. *AnonymAI* is a nine months research project co-financed by the H2020 project "NGI Trust"[28]. *AnonymAI*'s project goal consists in providing a legally compliant solution for the anonymization of textual documents. The adopted approach allows for an integration between innovative technologies and regulatory obligations, making it possible to specify anonymization profiles that are at the same time legally compliant and customized to the needs of end users (i.e., avoiding to mask data that can be relevant for secondary linguistic analysis whenever possible).

The anonymization profile applied to EVALITA4ELG dataset detects and masks the following direct and indirect identifiers:

- `Person Names` masked as PERSON
- `Phone Numbers` masked as PHONE
- `Emails` masked as EMAIL
- `Mentions/Reply/Retweet` masked as MENTION
- `URLs` masked as URL

The labels used to mask the personal information that we want to protect are enriched with a number in order to distinguish between different entities of the same kind mentioned within the same text (e.g., PERSON_1 told PERSON_2 that PERSON_3 was going to arrive). The tool was configured in order to exclude public person names (as politicians, journalists, actors) from the anonymization process, since they are particularly relevant in the context of proposed evaluation tasks (e.g., SardiStance (Cignarella et al. 2018) or Sentipolc (Basile et al. 2018), focusing on stance on political debates or on sentiment in political domain), as well as because public person names are not included in data protection procedures. Public persons to be excluded from anonymization were specified in a black list, manually compiled by reviewing the list of candidate person names extracted in a preliminary run of the anonymization tool.

The most frequent entities that were masked in the anonymization process consist of person names and mentions (e.g., in the SardiStance data set about 50 person names and 150 mentions).

5. The EVALITA Knowledge Graph

In this section we describe the first release of the EVALITA ontology[29], a knowledge graph that provides the essential information about the editions of the EVALITA evaluation campaign, which is described in terms of organized tasks, but also of people and institutions that constitute the community of EVALITA in its evolution over the years.

28 Grant agreement n.825618.
29 We use the terms *knowledge graph* and *ontology* synonymously in this paper.

The ontology is implemented in OWL and it will be soon available both on the website of the EVALITA4ELG project[30] and as a service on the ELG platform.

As namespace for our ontology we rely on the standard rdf namespace[31] as well as our own e4e namespace[32].

The current version of the ontology comprises 148 classes, 37 object properties and 9 data properties that describe the different editions of the EVALITA campaigns, the tasks organized therein and the systems submitted. The primary classes are:

- `EvaluationCampaign`: each edition of an evaluation campaign, such as *EVALITA 2020*, is represented as class instance. We use subclasses for different editions (e.g. EvaluationCampaign → EVALITA → EVALITA2020) to allow the future incorporation of data from other evaluation campaigns such as Semeval or IberLEF.
- `Task`: instances correspond to the different tasks organised in each edition. Some of them have subtasks (e.g. the main task corresponds to the detection of hate speech instances; the subtask is classifying the kind of hate speech). While tasks are modelled as classes, their subtasks are modelled as subclasses of the respective classes.
- `Person`: information (e.g. name and surname) about researchers involved in the organisation or participation in a task.
- `Institution`: universities, research centres or companies involved in the organisation or participation in a task.

The primary classes are linked to each other trough a set of relations[33]: (i) the `hasOrganizer` property links a task to its organizers, i.e. instances in the `Person` class; (ii) similarly, `hasParticipant` relates a task to the participants that submitted a system; (iii) `hasSystem` models the relation between a task and the systems submitted; (iv) an evaluation campaign is related to the tasks/subtasks organized therein through the `hasTask / hasSubTask` property[34]; (v) `hasInstitution` relates a task or a system to the institution its organizers belong to. We preferred to relate the institution to the task, rather than to the class `Person`. The reason behind this choice is that a researcher may have changed its affiliation through time, thus resulting in multiple affiliations for a single person, without the possibility to understand which task he/she organized while working at a certain institution. Our modelling choice allows us to observe the presence of different research centres in the organization or participation to a task, and the evolution through the different editions.

The ontology organizes the available information on tasks through various properties that link a `Task` to the following classes:

- `Dataset`: the name of the resource (corpus or lexical conceptual resource) used in the task for training and evaluation;

30 `http://evalita4elg.di.unito.it`
31 `http://www.w3.org/1999/02/22-rdf-syntax-ns`
32 The ontology and along with it the e4e namespace will be made available as a subdomain of our project website which is located at: `http://evalita4elg.di.unito.it`
33 We use the terms *relation* and *Object Properties* interchangeably.
34 In our implementation, the former relations are all subsumed by a super-relation called `hasPart`.

- `Language`: the language considered;
- `System`: the name of the systems that participated in the task;
- `NLPTask`, `NLPTaskDomain`, `ComputationalTask`: features defining the type of task.

The task is thus related to its dataset through the property `hasDataset`, the relation `hasLanguage` specifies the language, and `hasSystem` relates task to the systems submitted therein.

The characteristics of the dataset are described as classes, specifically:

- `MediaType`: it indicates the kind of data collected in the resource; possible values are text, audio, image, video, ngram. It is specified through the property `hasMediaType`, that has the class `Dataset` as domain and `MediaType` as range.
- `DatasetFormat`: the format of the resource, e.g. xml, csv, json. It is expressed through the object property `hasDatasetFormat`.
- `DatasetGenre`: if the resource is a corpus, we specify the genre of the texts collected (newspaper articles, reviews, social media texts, etc.), and link it to the dataset by means of the relation `hasDatasetGenre`.
- `License`: the licence or terms of use with which the resource is distributed, specified by the relation `hasLicense`
- `_Annotation_`: each task or subtask and the corresponding dataset are linked to an annotation element, which is a description of the annotation schema used in a dataset. Since it is not possible to describe a property with more than two arguments (RDF object and data properties are binary indeed, respectively linking two individuals or an individual with a value), we represent their annotation as a class rather than as a property, i.e. a reified relation. Thus the class `_Annotation_` allows us to describe for each annotation individual the unit of annotation considered (`AnnotationUnit`, e.g. word, word sequence, sentence, document), the type of annotation (`AnnotationType`, e.g. categorical, multilabel, continuous), and the context considered during the annotation (`AnnotationContext`). Therefore, an instance of the `_Annotation_` class represent a specific annotation scheme.

For each task, the teams that participated in the competition are listed, grouped in the class `Team`. A team has the properties `hasParticipant`, that describes the researchers that composed it (listed in the above mentioned class `Person`), `hasInstitution` (as for the `Task` individuals) and `hasDeveloped`, that describe the systems developed by the team, grouped in the class `System`. Teams and systems are also linked to the edition of the evaluation campaign and the task they took part in, through the property `isParticipantOf`.

Systems are further described for the class of algorithms used (property: `hasComputationalMethod`, class `ComputationalMethod`) and their performance in each subtasks. As for annotations, also systems performance is represented through a reified relation, i.e. the class `_Evaluation_`. It allows us to describe for each evaluation individual the system it refers to, the subtask the system partici-

pated in, the evaluation measure used (e.g. accuracy, F1 score, listed in the class `EvaluationMeasure` and linked to the evaluation individual through the property `hasEvaluationMeasure`), and the performance score of the system (as a numerical data property, `hasPerformance`).

It is worth pointing out that most of the object properties in the knowledge graph have an inverse property related. For example, the above mentioned relation `hasDataset` - that has as domain an individual of the class `Task` and as range a dataset - is linked to its inverse property `isDatasetOf` that has as domain a dataset and as range a task. Thus, it is possible to search in the knowledge graph not only the dataset used for a specific task, but also the name of the task a known dataset was used for. Furthermore, we have used the chaining mechanism to transfer properties where it was plausible. For instance, if a task t has EvaluationMeasure e, then all systems that participated in t also have e as EvaluationMeasure.

Knowledge Graph Visualization. Figure 2 depicts the structure of our ontology in accord with the previous descriptions. The primary classes `EvaluationCampaign`, `Task`, `Institution` and `Person` are color-coded and we show how they are related to each other on the instance of the *HaSpeeDe2018* task.

We observe in the figure that *HaSpeeDe2018* was a task in the *Evalita2018* evaluation campaign as defined by the `isTaskOf` relation. The `hasOrganizer` relation links the task to its five organizers.[35] Analogously, the institutions that were involved in the task are specified via the `hasInstitution` relation. Besides the previously described object properties, we specified a number of data properties, of which two are illustrated. First, the website of the task is specified via the `hasWebsite` relation, and secondly, the technical paper describing the task is connected via the `hasPaper` relation. Finally, note how the specific task of *HaSpeeDe2020*, also displayed in Figure 2 is a re-run of the *HaSpeeDe2018* task, in which novel systems competed to solve a very similar task on a very similar dataset. This is specified by the `isReRunOf` relation from the 2020 task to the 2018 task.

In Figure 3 we take a different perspective by focusing on a specific evaluation campaign, *Evalita2018*. The different tasks that took place in the campaign are connected via the `hasTask` relation. The tasks themselves are related to their institutions and organizers, as previously illustrated, via the respective relations.

5.1 Knowledge Graph Querying

The knowledge graph can be queried either through the Protégé interface or through a SPARQL endpoint, that will be soon made available to the public. In this section we show some examples of SPARQL queries, with a focus on the researchers and institutions involved in the organization of the tasks.

The SPARQL language allows to inspect the ontology selecting some variables that occur among the set of triples (subject, predicate, object) composing the knowledge graph. It is thus possible to answer relevant questions related to the EVALITA campaign, extracting information from our ontology. For instance, *what is the total number of institutions involved as organizers of tasks in all seven EVALITA campaigns?*. We retrieve the

35 Note, that we collect all organizers here, but in the knowledge graph there are five relations between the task and each of the organizers.

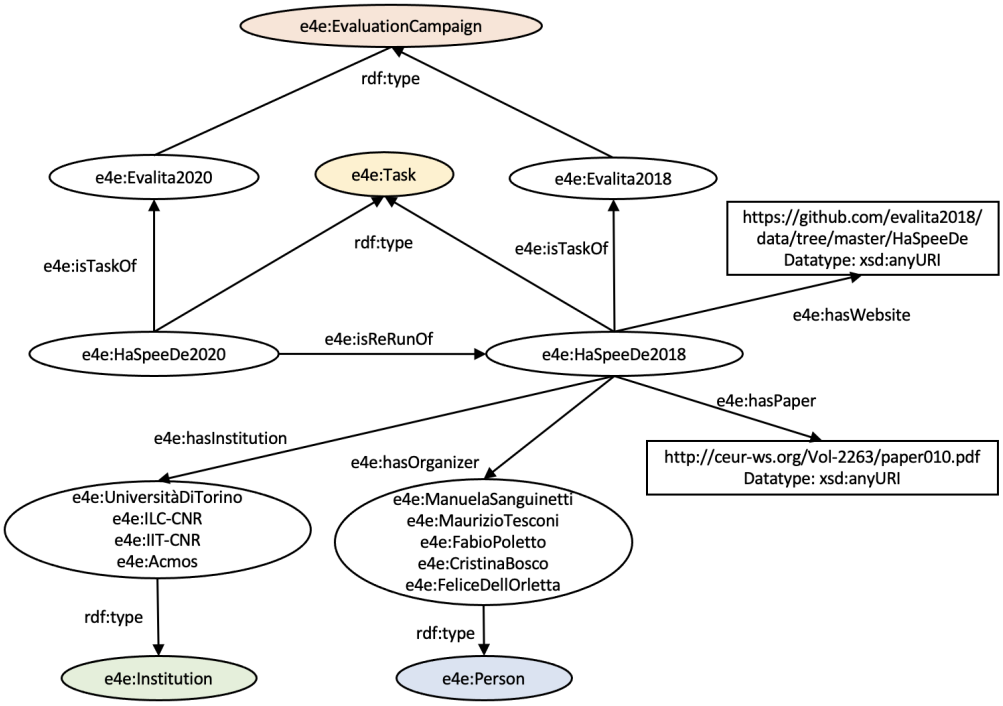

Figure 2
Illustration of the structure of the knowledge graph. Primary classes are colored and their relations are illustrated around the HaSpeeDe2018 task.

relevant triples (task, hasInstitution, institution) as follows and find the result to be 55 distinct institutions involved:

```
SELECT  (COUNT(distinct ?institution) AS ?totalInstitutions)
where {
  ?task e4e:hasInstitution ?institution.
}
>>>> result: 55 <<<<
```

We ask more specifically *how did the number of institutions per EVALITA edition change over time?*:

```
SELECT ?edition (COUNT(distinct ?institution) AS ?totalInstitutions)
where {
  ?task e4e:hasInstitution ?institution.
  ?task e4e:isTaskOf ?edition.
}
```

The result is depicted in Figure 4. It shows the overall rising number of institutions over time which are involved in the different EVALITA editions. In the first edition of EVALITA, only six institutions were involved in the organization of the six tasks, one for each task, with the exception of that focused on parsing where researchers from the University of Torino and the University of Roma Tor Vergata collaborated. However,

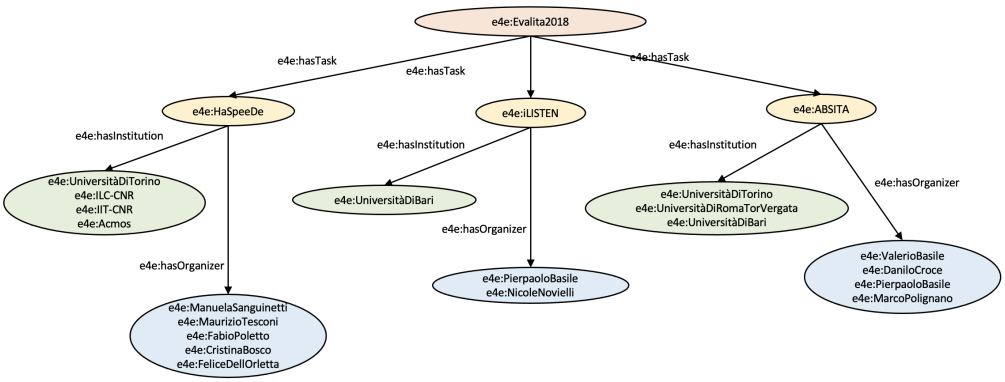

Figure 3
Closer look at the instantiation of an `EvaluationCampaign` and its surrounding relations. Example for the *Evalita2018* campaign.

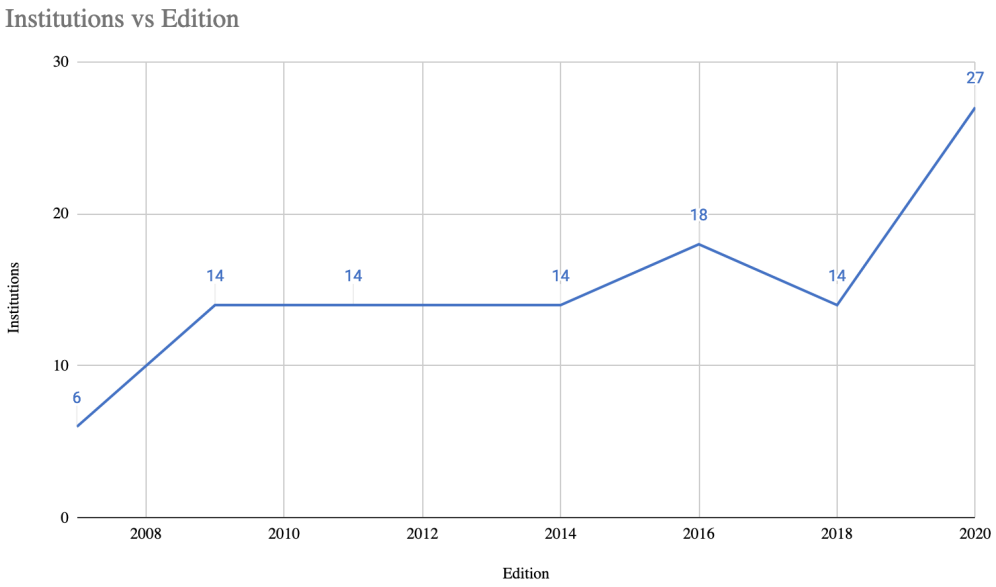

Figure 4
Number of institutions involved as task organizers in the different editions of EVALITA.

this number has progressively grown starting from the second edition and has kept almost stable until the latest EVALITA, where the higher amount of participants and organized tasks both has been achieved: 28 different organizations participated and 14 tasks were organized. Moreover, if we look at the number of institutions per task and average them by edition, we observe that from the second edition the collaboration among institutions has grown (from an average of 1.2 different institutions in 2007 to 2.1 in 2009), with a peak in the 2016 and 2020 editions (with average number of institutions per task 3.5 and 3.3 respectively, see Table 5.1). The peak of the 2016 edition could

Table 1
Number of institutions involved as organizers in each task in the different editions of EVALITA.

EVALITA edition	Number of tasks organized by n institutions						Average institutions per task
	n=1	n=2	n=3	n=4	n=5	n=6	
2007	4	1					1.2
2009	3	2	2	1			2.1
2011	4	5	1	1			1.9
2014	2	2	2	2	1		2.8
2016		1	3	1		1	3.5
2018	3	3	3	1			2.2
2020		4	4	4	2		3.3

be related to the mentioned effort to encourage task organisers to collaborate on the creation of a shared test set across tasks, with the aim of providing a unified framework for multiple layers of annotations related to different phenomena over exactly the same social media data, so as to facilitate and promote the development and testing of end-to-end systems (Basile et al. 2017).

We can further look at the first year of participation of an institution as task organizer with the following query, which can give us an overview of the growth of the Italian NLP community, showing historical seats and new members:

```
SELECT ?inst (min(?edition) as ?firsted)
where {
  ?task e4e:hasInstitution ?inst.
  ?task e4e:isTaskOf ?edition.
}
group by ?inst
order by ?edition
```

We show the first ten items of the result of this query in table5.1.

As a last example query, we can inspect the most active institutions as organizers, whose first ten items are listed in Table 5.1:

```
SELECT ?inst (COUNT(distinct ?task) AS ?totaltask)
where {
  ?task e4e:hasInstitution ?inst.
}
group by ?inst
order by (-?totaltask)
```

5.2 EVALITA Resources and Community: an Ontology-Driven Survey

In section 5.1, we have shown some examples on how SPARQL queries can be exploited to extract data and information from the EVALITA knowledge graph, gaining a historical perspective on the campaign. In this section, taking the same approach, we will

Table 2
First year of participation as task organizer of institutions (first ten items of the query result).

Institution	First edition of participation as task organizer
CELCT	Evalita2007
FBK	Evalita2007
UniversitàDiBologna	Evalita2007
UniversitàDiRomaTorVergata	Evalita2007
UniversitàDiTorino	Evalita2007
ABLA	Evalita2009
ILC-CNR	Evalita2009
Loquendo	Evalita2009
SaarlandUniversity	Evalita2009
Speechcycle	Evalita2009

Table 3
First ten most active institutions and number of tasks organized.

Institution	Number of tasks organized
Università di Torino	16
ILC-CNR	14
FBK	10
Università di Bari	9
Università di Pisa	9
Università di Napoli Federico II	8
University of Groningen	7
Università di Bologna	7
Università di Trento	7
Universitat Politècnica de València	5

systematically rely on the availability of the ontology, in order to extract information useful to draw some general considerations, on the one hand, about the resources developed and used in the different EVALITA tasks, on the other hand, about people involved over the years in the EVALITA initiatives, with different roles as participants and organizers. This survey aims at providing an overview of the data sets made available as benchmarks to the Italian NLP community and so highlighting some general trends about the typology of data that have been of interest through the EVALITA editions, and the annotations produced on the data, to make explicit the relevant knowledge for training and testing different NLP models, challenged on various increasingly complex and fine-grained tasks.

The EVALITA campaign has reached its seventh edition in 2020, with a total of 62 organized tasks. As highlighted by (Basile et al. 2020) the various editions have shown an increasing involvement of participants, starting from 30 in 2007. Attendance doubled in 2016 (60 participants), then has been attested in the last editions on numbers that

overcome the hundred researchers, with 115 participants in 2018 and the record number of 130 participants in 2020.

Similarly, the number of people involved as task organizers has growth from 11 in 2007 to 59 in 2020. These numbers reflect the growth of the Italian NLP community, as well as the success of the evaluation campaign. The type of tasks also evolved through time: from more traditional challenges, like parsing or PoS-tagging, the interest of the community shifted in the recent years to new challenges, with a growing interest in semantics, pragmatics and affect.

For building the evaluation exercises where systems performance were compared, the organizers have faced the need for suitable benchmarks, whose creation and annotation in many cases required meaningful efforts. The Italian NLP community can now profit of this work that has widened the set of resources available for the Italian language. During the different editions, indeed, the kind of data sets elaborated evolved together with the type of tasks organized, bringing new types of language data or new layers of annotation, applied also in a cross-task perspective, and leading to the proposal of interesting evaluation settings designed to test adaptability of systems to different textual domains and genres, or aspects related the diachronic dimension, see, e.g., shared tasks in the 2020 “Time and Diachrony" track, DIACR-Ita and DaDoEval.

In the first editions, tasks have mainly used balanced or general reference corpora. It has been the case in 2007, where the CORIS/CODIS corpus, the TUT treebank, the ISST treebank, and the I-CAB corpus were employed. All these resources share the characteristic of being corpora of the written language, frequently including newspaper articles or being a balanced sample of different genres. The CORIS/CODIS corpus (Rossini Favretti, Tamburini, and De Santis 2002), released in 2001, is a synchronic corpus of written language, with texts of different genres written between the 1980s and 1990s. The I-CAB corpus (Magnini et al. 2006), used both in the TERN (Bartalesi Lenzi and Sprugnoli 2007) and NER tasks (Speranza 2007), is constituted by texts from the *L'Adige* newspaper, annotated with semantic information, i.e. temporal expressions, named entities and relations between entities. Those annotations were exploited in the tasks the corpus was used for. Similar textual genres constitute the TUT treebank and the ISST treebank. TUT (Turin University Treebank, (Bosco et al. 2000)) comprehends multiple genres, but for the 2007 parsing task (Bosco, Mazzei, and Lombardo 2007) two equally sized subcorpora were used, one from the Italian legal Code and one from Italian newspapers. ISST (Italian Syntactic Semantic Treebank, (Montemagni et al. 2003)) includes texts from newspapers, periodicals and financial articles, but for the 2007 word sense disambiguation task (Bertagna, Toral, and Calzolari 2007) only the newspaper section was used.

Since 2009, data from computer-mediated communications started being used. Wikipedia articles were collected to prepare the dataset for the POS tagging and the textual entailment task (Attardi and Simi 2009; Bos, Zanzotto, and Pennacchiotti 2009) in 2009, whereas in 2011 this genre was used in the parsing and super sense tagging challenges (Bosco and Mazzei 2012a, 2012b; Dei Rossi, Di Pietro, and Simi 2013). From 2014, a specific genre of computer-mediated language was introduced: social media data. Specifically, in the SENTIPOLC 2014 task (Basile et al. 2014, 2018), Italian texts from the social networks Twitter were used in a sentiment classification challenge. In the following editions of EVALITA the use of this genre of text spread out: 4 tasks in 2016 (Minard et al. 2016; Basile et al. 2016; Bosco et al. 2016; Barbieri et al. 2016), 5 in 2018 (Fersini, Nozza, and Rosso 2018; Dell’Orletta and Nissim 2018; Bosco et al. 2018; Ronzano et al. 2018; Cignarella et al. 2018), 5 in 2020 (Fersini, Nozza, and Rosso 2020; Brunato et al. 2020; Miliani et al. 2020; Sanguinetti et al. 2020; Cignarella et al. 2020). In

most cases, data were extracted from Twitter, probably because of its favorable policies on data access and use, and clear terms of services, but in some cases other social networks were used as source, i.e., Facebook (like in HaSpeeDe 2018), youtube comments (GxG 2018 (Dell'Orletta and Nissim 2018)), Instagram for memes (DankMemes 2020, (Miliani et al. 2020).

Other less frequently employed types of linguistic data are transcribed spoken language data (NERTBN 2011, iLISTEN 2018, KIPoS 2020) and not naturalistic data, i.e., made up texts like those used in the AcCompl-it task (Brunato et al. 2020). A peculiar case of data are the datasets used for the NLP4FUN (Basile et al. 2018) and Ghigliottin-AI (Basile et al. 2020) tasks: here the dataset was constituted by examples of a popular language game, where the aim is to guess a word linked to other 5 words by various relations.

The varieties of textual genres just mentioned has been exploited in some tasks to test the systems adaptation to different domains. Frequently indeed, the performance of a system in a domain different from the trained one decrease. Different tasks have proposed challenges that check for the flexibility of the systems. Among such tasks, let us mention: 1) parsing task evaluation across domains (legal, newspaper), 2) HaSpeeDe evaluation across domains (Twitter, Facebook, headlines), 3) KIPOS evaluation across domains (formal/informal spoken language), 4) GxG evaluation on author profiling in terms of gender across various genres (Twitter, YouTube, Children writing, News/journalism, Personal diaries).

The attention in EVALITA to the performance of systems among different domains and genres is confirmed by the fact that some traditional tasks have been reproposed during the various editions of EVALITA with a focus on different textual data over the years. It is the case for example of the POS tagging task. In 2007 (Tamburini 2007) it was mainly restricted to newspaper and literature data; in 2009 (Attardi and Simi 2009), it was expanded with domain adaptation between the training and the test set: systems were trained on newspaper articles from La Repubblica corpus (Baroni et al. 2004), and then evaluated on Wikipedia texts. In the 2016 task PosTwita (Bosco et al. 2016), the challenge of POS tagging has been brought in the social media language domain, asking participants to adapt their tagging systems to this particular text domain. The dataset comprehends tweets that were also annotated for sentiment (Basile et al. 2014, 2016). Lastly, in 2020, this challenge was applied to the spoken language (Bosco et al. 2020), employing the KiParla corpus (Mauri et al. 2019).

Among the different editions of parsing tasks (2007, 2009, 2011), the dataset used changed both in terms of genres and annotations provided. With respect to the 2007 edition (Bosco, Mazzei, and Lombardo 2007), the second edition (Bosco et al. 2009) proposed the task using two different kinds of annotations - they used both constituency and dependency-based annotations; the third edition (Bosco and Mazzei 2012a, 2012b) added new genres (specifically, legal and Wikipedia texts) to the challenge. Similarly, the third edition of the named entity recognition task (Bartalesi Lenzi, Speranza, and Sprugnoli 2013), provided a new corpus of transcribed news broadcast for systems evaluation, moving from previous editions where newspaper corpora were used.

In 2016, four tasks (Barbieri et al. 2016; Basile et al. 2016; Minard et al. 2016; Bosco et al. 2016) used at least part of the same dataset (mainly taken from the dataset used in SENTIPOLC 2014), providing different layers of annotation for the same data (Basile et al. 2017).

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, memes. However, in 2014, speech tasks have been more numerous than textual ones (4 our of 7), whereas in the first and latest editions there were no tasks that focused on speech data. The speech tasks have been made possible thanks to the participation of members of the AISV[36] community in the campaign, starting from the 2009 edition.

The speech tasks promoted various challenges in the community, namely: identification of the speaker (Aversano, Brümmer, and Falcone 2009, HDMI); identification of the language/dialect (Romano and Russo 2014), the evaluation of spoken dialogue systems (Cutugno et al. 2018, IDIAL); detection, transcription or recognition of speech (Coro, Gretter, and Matassoni 2009, Connected digits recognition); automatic speech recognition (Matassoni, Brugnara, and Gretter 2013); speech recognition at different speed rate (Badino 2016, ArtiPhon); speech recognition oriented at domestic robotics (Cutugno et al. 2018, SUGAR), (Brutti, Ravanelli, and Omologo 2014, SASLODOM); alignment of audio sequences to the relative transcriptions (Cutugno, Origlia, and Seppi 2013, FACS); , emotion recognition in speech (Origlia and Galata 2014, ERT).

With regard to the language, in most cases the task and the data set concerned the Italian language. However, a few tasks comprehended the evaluation for other languages in addition to Italian, in a multilingual setting. In 2014 the HDMI task (Romano and Russo 2014) challenged participants to build systems that recognise the language of speech input among 18 languages and 20 dialects. In 2018 the AMI task (Fersini, Nozza, and Rosso 2018) was aimed at recognizing and categorizing misogynistic tweets in both Italian and English samples. In 2020 the CONcreTEXT challenge (Gregori et al. 2020) asked systems to automatically assign a concreteness score to words in context both for English and Italian. The inclusion of a parallel evaluation for another language may have fostered the participation of teams for foreign institutions: in AMI, for example, most teams were composed by researchers from foreign institutions, even if some of them participated only in the English subtask. A multilingual setting can make the EVALITA campaign known among researchers in other countries and create other opportunities to create link between the Italian and the international NLP communities.

6. Conclusion

This paper describes an ongoing project whose main aim is at extending the coverage of the ELG platform to the resources developed within the context of the evaluation campaigns EVALITA. The steps of the project are surveyed showing the relevant contribution they can offer to the research community working on NLP and especially on Italian language. The project outcomes also include the survey of the material developed in the campaigns, its organization in an ontology that makes accessing them easier, the application of procedures for making them compliant with current standards for data sharing.

Besides creating a complete and accessible catalogue of benchmarking resources and NLP models and tools for the Italian language, contributing to widening the coverage of the Italian LTs in the ELG platform, EVALITA4ELG project includes as further research goal the creation of a unified benchmark for Italian Natural Language Understanding, similar to the GLUE (Wang et al. 2018) and SuperGLUE (Wang et al. 2019) initiatives. We start from an advantageous position, for two reasons: we already

36 Associazione Italiana Scienze della Voce: `https://www.aisv.it/`

collected and catalogued a large number of benchmarks, including data and evaluation metrics and scripts, and the ELG provides a stable technological infrastructure to host a Web-based platform to run the evaluation and keep track of the scores publicly.

Finally, we observe that the interest in testing portability of models across tasks and languages is growing steadily, as well as the interest in experimenting with cross- and multi-task learning, which makes the development of multilingual benchmarks one of the new frontiers.

Acknowledgments

The EVALITA4ELG project was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

References

AA.VV. 2009. Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence. volume 9, Reggio Emilia, Italy, December.

Attardi, Giuseppe, 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*, 9:43–61.

Attardi, Giuseppe and Maria Simi. 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.

Aversano, Guido, Niko Brümmer, and Mauro Falcone. 2009. EVALITA 2009 Speaker Identity Verification Application Track-Organizer's Report. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, volume 9, Reggio Emilia, Italy, December.

Badino, Leonardo. 2016. The ArtiPhon Task at Evalita 2016. In *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)*, volume 1749, Napoli, December. CEUR Workshop Proceedings (CEUR-WS. org).

Barbieri, Francesco, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 sentiment polarity classification task. In *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)*, volume 1749, Napoli, December. CEUR Workshop Proceedings (CEUR-WS. org).

Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI (XML)- compliant Corpus of Newspaper Italian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1771–1774, Lisbon, May. European Language Resources Association (ELRA).

Bartalesi Lenzi, Valentina, Manuela Speranza, and Rachele Sprugnoli. 2013. Named Entity Recognition on Transcribed Broadcast News at EVALITA 2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian. International Workshop, EVALITA 2011, Revised Selected Papers*, pages 86–97, Rome, January 24-25, 2012. Springer.

Bartalesi Lenzi, Valentina and Rachele Sprugnoli. 2007. Description and Results of the TERN Task. *Intelligenza Artificiale*, 4(2):55–57, June.

Basile, Pierpaolo, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in ITalian tweets (NEEL-IT) task. In *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)*, volume 1749, Napoli, December. CEUR Workshop Proceedings (CEUR-WS. org).

Basile, Pierpaolo, Marco de Gemmis, Lucia Siciliani, and Giovanni Semeraro. 2018. Overview of the EVALITA 2018 Solving Language Games (NLP4FUN) Task. In Tommaso Caselli, Nicole

Novielli, Viviana Patti, and Paolo Rosso, 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, Pierpaolo, Marco Lovetere, Johanna Monti, Antonio Pascucci, Federico Sangati, and Lucia Siciliani. 2020. Ghigliottin-AI@EVALITA2020: Evaluating Artificial Players for the Language Game "La Ghigliottina" (short paper). In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *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, December. CEUR Workshop Proceedings (CEUR-WS. org).

Basile, Pierpaolo, Malvina Nissim, Rachele Sprugnoli, Viviana Patti, and Francesco Cutugno. 2017. EVALITA Goes Social: Tasks, Data, and Community at the 2016 Edition. *Italian Journal of Computational Linguistics*, 3(1):93–127.

Basile, Valerio, Andrea Bolioli, Viviana Patti, Paolo Rosso, and Malvina Nissim. 2014. Overview of the EVALITA 2014 SENTIment POLarity Classification task. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, 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, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics (ACL).

Basile, Valerio, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 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)*, volume 2765, Online event, December 17th, 2020. CEUR Workshop Proceedings (CEUR-WS. org).

Basile, Valerio, Nicole Novielli, Danilo Croce, Francesco Barbieri, Malvina Nissim, and Viviana Patti. 2018. Sentiment Polarity Classification at EVALITA: Lessons Learned and Open Challenges. *IEEE Transactions on Affective Computing*, 12(2):466–478.

Bertagna, Francesca, Antonio Toral, and Nicoletta Calzolari. 2007. EVALITA 2007: The All-Words WSD Task. *Intelligenza Artificiale*, IV(2):50–52, June.

Bos, Johan, Fabio Massimo Zanzotto, and Marco Pennacchiotti. 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, Cristina, 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, December. CEUR Workshop Proceedings (CEUR-WS. org).

Bosco, Cristina, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, 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).

Bosco, Cristina, Tamburini Fabio, Bolioli Andrea, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 Part Of Speech on Twitter for ITAlian task. In *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)*, volume 1749, Napoli, December. CEUR Workshop Proceedings (CEUR-WS. org).

Bosco, Cristina, Vincenzo Lombardo, Leonardo Lesmo, and Vassallo Daniela. 2000. Building a treebank for italian: a data-driven annotation schema. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 99–105, Athens, May - June. European Language Resources Association (ELRA).

Bosco, Cristina and Alessandro Mazzei. 2012a. The Evalita 2011 parsing task: the constituency track. In *Working Notes of Evalita 2011*, Roma, January. CELCT a r.l.

Bosco, Cristina and Alessandro Mazzei. 2012b. The Evalita 2011 parsing task: the dependency track. In *Working Notes of Evalita 2011*, Roma, January. CELCT a r.l.
Bosco, Cristina, Alessandro Mazzei, and Vincenzo Lombardo. 2007. An analysis of the first parsing system contest for italian. *Intelligenza Artificiale*, IV(2):30–33, June.
Bosco, Cristina, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice Dell'Orletta, and Alessandro Lenci. 2009. Evalita'09 parsing task: comparing dependency parsers and treebanks. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, volume 9, Reggio Emilia, Italy, December.
Brunato, Dominique, 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, December. CEUR Workshop Proceedings (CEUR-WS. org).
Brutti, Alessio, Mirco Ravanelli, and Maurizio Omologo. 2014. SASLODOM: Speech Activity detection and Speaker LOcalization in DOMestic environments. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 139–146, Pisa, Italy, December. Pisa University Press.
Caselli, Tommaso, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Evalita 2018: Overview on 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 the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Torino, December. CEUR Workshop Proceedings (CEUR-WS. org).
Cignarella, Alessandra Teresa, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on Irony Detection in Italian Tweets (IRONITA). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, 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).
Cignarella, Alessandra Teresa, 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, December. CEUR Workshop Proceedings (CEUR-WS. org).
Coro, Gianpaolo, Roberto Gretter, and Marco Matassoni. 2009. Evalita 2009: Description and results of the speech recognition task. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, volume 9, Reggio Emilia, Italy, December.
Cutugno, Francesco, Maria Di Maro, Sara Falcone, Marco Guerini, Bernardo Magnini, and Antonio Origlia. 2018. Overview of the Evalita 2018 Evaluation of Italian Dialogue Systems (IDIAL) task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, 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).
Cutugno, Francesco, Antonio Origlia, and Dino Seppi. 2013. Evalita 2011: Forced alignment task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian.International Workshop, EVALITA 2011, Revised Selected Papers*, pages 305–311, Rome, January 24-25, 2012. Springer.
Dei Rossi, Stefano, Giulia Di Pietro, and Maria Simi. 2013. Description and Results of the SuperSense Tagging Task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian.International Workshop, EVALITA 2011, Revised Selected Papers*, pages 166–175, Rome, January 24-25, 2012. Springer.
Dell'Orletta, Felice and Malvina Nissim. 2018. Overview of the EVALITA 2018 cross-genre gender prediction (gxg) task. In *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).

Dell'Orletta, Felice and Malvina Nissim. 2018. Overview of the Evalita 2018 cross-genre gender prediction (GxG) task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, 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).

Fersini, Elisabetta, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, 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).

Fersini, Elisabetta, 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, December. CEUR Workshop Proceedings (CEUR-WS. org).

Gregori, Lorenzo, 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, December. CEUR Workshop Proceedings (CEUR-WS. org).

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics (ACL).

Labropoulou, Penny, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez-Pérez, and Andres Garcia-Silva. 2020. Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3428–3437, Marseille, May. European Language Resources Association (ELRA).

Magnini, Bernardo, Amedeo Cappelli, Fabio Tamburini, Cristina Bosco, Alessandro Mazzei, Vincenzo Lombardo, Francesca Bertagna, Nicoletta Calzolari, Antonio Toral, Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Manuela Speranza. 2008. Evaluation of natural language tools for Italian: EVALITA 2007. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2536–2543, Marrakech, May. European Language Resources Association (ELRA).

Magnini, Bernardo, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors. 2013. *Evaluation of Natural Language and Speech Tools for Italian, International Workshop, EVALITA 2011, Revised Selected Papers*, volume 7689 of *Lecture Notes in Computer Science*, Rome, Italy, January 24-25, 2012. Springer.

Magnini, Bernardo, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 963–968, Genova, May. European Language Resources Association (ELRA).

Matassoni, Marco, Fabio Brugnara, and Roberto Gretter. 2013. Evalita 2011: Automatic speech recognition large vocabulary transcription. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian.International Workshop, EVALITA 2011, Revised Selected Papers*, pages 274–285, Rome, January 24-25, 2012. Springer.

Mauri, Caterina, Silvia Ballarè, Eugenio Goria, Massimo Cerruti, and Francesco Suriano. 2019. KIParla Corpus: A New Resource for Spoken Italian. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481, Bari, November. CEUR Workshop Proceedings (CEUR-WS. org).

Miliani, Martina, 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, December. CEUR Workshop Proceedings (CEUR-WS. org).

Minard, Anne-Lyse, Manuela Speranza, Tommaso Caselli, and Fondazione Bruno Kessler. 2016. The EVALITA 2016 event factuality annotation task (FactA). In *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)*, volume 1749, Napoli, December. CEUR Workshop Proceedings (CEUR-WS. org).

Montemagni, Simonetta, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003. Building the Italian Syntactic-Semantic treebank. In Anne Abeillé, editor, *Treebanks. Building and Using Parsed Corpora*. Springer, pages 189–210.

Nissim, Malvina, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. 2017. Last Words: Sharing Is Caring: The Future of Shared Tasks. *Computational Linguistics*, 43(4):897–904, December.

Origlia, Antonio and Vincenzo Galata. 2014. EVALITA 2014: Emotion Recognition Task (ERT). In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 112–115, Pisa, Italy, December. Pisa University Press.

Proisl, Thomas and Gabriella Lapesa. 2020. KLUMSy@ KIPoS: Experiments on Part-of-Speech Tagging of Spoken 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, December. CEUR Workshop Proceedings (CEUR-WS. org).

Rangel, Francisco and Paolo Rosso. 2018. On the implications of the general data protection regulation on the organisation of evaluation tasks. *Language and Law*, 5(2):95–117.

Rehm, Georg, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajič, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiljevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez-Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. 2020a. European Language Grid: An Overview. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3366–3380, Marseille, May. European Language Resources Association (ELRA).

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajič, Khalid Choukri, Andrejs Vasiljevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez-Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadiņa, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon. 2020b. The European language technology landscape in 2020: Language-centric and human-centric AI for cross-cultural communication in multilingual Europe. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3322–3332, Marseille, May. European Language Resources Association (ELRA).

Rehm, Georg and Hans Uszkoreit. 2013. *META-NET strategic research agenda for multilingual Europe 2020*. Springer.

Romano, Antonio and Claudio Russo. 2014. Human and Machine Language/Dialect Identification from Natural Speech and Artificial Stimuli: a Pilot Study with Italian Listener. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 131–138, Pisa, Italy.

Pisa University Press.
Ronzano, Francesco, Francesco Barbieri, Endang Wahyu Pamungkas, Viviana Patti, Francesca Chiusaroli, et al. 2018. Overview of the Evalita 2018 Italian emoji prediction (ITAMOJI) task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, 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).
Rossini Favretti, Rema, Fabio Tamburini, and Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Andrew Wilson, Paul Rayson, and Tony McEnery, editors, *A rainbow of corpora: Corpus linguistics and the languages of the world*. Lincom-Europa, Munich, pages 27–38.
Sanguinetti, Manuela, 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, December. CEUR Workshop Proceedings (CEUR-WS. org).
Speranza, Manuela. 2007. The Named Entity Recognition task. *Intelligenza Artificiale*, IV(2), June.
Sprugnoli, Rachele, Viviana Patti, and Franco Cutugno. 2016. Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, 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)*, volume 1749, Napoli, December. CEUR Workshop Proceedings (CEUR-WS. org).
Tamburini, Fabio. 2007. The part-of-speech tagging task. *Intelligenza Artificiale*, IV, June.
Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.
Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, November. Association for Computational Linguistics (ACL).