GRIT: A Dataset of Group Reference Recognition in Italian

Sergio E. Zanotto[1], **Qi Yu**[1], **Miriam Butt**[1], **Diego Frassinelli**[1,2]
[1]Department of Linguistics & Cluster of Excellence "The Politics of Inequality", University of Konstanz
[2]Center for Information and Language Processing, LMU Munich
{sergio.zanotto, qi.yu, miriam.butt, diego.frassinelli}@uni-konstanz.de

Abstract

For the analysis of political discourse a reliable identification of group references, i.e., linguistic components that refer to individuals or groups of people, is useful. However, the task of automatically recognizing group references has not yet gained much attention within NLP. To address this gap, we introduce GRIT (**G**roup **R**eference for **It**alian), a large-scale, multi-domain manually annotated dataset for *group reference recognition* in Italian. GRIT represents a new resource for automatic and generalizable recognition of group references. With this dataset, we aim to establish group reference recognition as a valid classification task, which extends the domain of Named Entity Recognition by expanding its focus to literal and figurative mentions of social groups. We verify the potential of achieving automated group reference recognition for Italian through an experiment employing a fine-tuned BERT model. Our experimental results substantiate the validity of the task, implying a huge potential for applying automated systems to multiple fields of analysis, such as political text or social media analysis.

Keywords: group reference recognition, NLP for social sciences, Italian language resource

1. Introduction

The rhetorical power of *group references*, i.e., linguistic expressions that literally or figuratively refer to individuals or groups of people, is of particular interest in the social sciences. For example, several studies have explored how group references serve as a tool to gain the favor of the addressed groups and consequently affect their voting behaviors (e.g., Strom, 1990; Wodak, 2012; Thau, 2019). When dealing with large-scale text analyses in this strand of research, the manual identification of group references is labor-intensive and time-consuming. There is thus a need for the automatic and reliable detection of group references. We name this task as **group reference recognition** (GRR henceforth).[1] A successful automatization of GRR has the potential of providing large quantity of empirical data for political communication studies, such as identifying the targets of political parties in social media data and party manifestos (Russmann, 2020; Horn et al., 2021). GRR presents a novel challenge as social group references can be expressed in very different ways. Consider the most representative classes of group references in (1)–(3), where the group references are underlined:

(1) **Proper nouns:**
[...] *the Zapatists were unarmed.*

(2) **Common nouns:**
The teachers and the students of [...]

(3) **Relative clauses:**
One of the reasons why the people who have no trust anymore [...]

Whereas group references using proper nouns as in (1) might be identified using existing named entity recognition (NER) tools, those using common nouns or relative clauses as in (2)–(3) are out of the scope of traditional NER. They require the identification of expressions that are not proper nouns, but still refer to groups of persons. Furthermore, group references are often figurative instead of literal: for example, the word 'Brazil' in 'Brazil won the World Cup' does not refer to the country, but is a metonymy referring to the Brazilian national football team. Even though the task of figurative language detection has been extensively studied (e.g., Teraoka, 2016; Gritta et al., 2017; Chakrabarty et al., 2021; Lai et al., 2023; Wang et al., 2023), there is no study on detecting figurative language usages that *specifically refer to social groups*.

GRR is not yet an established task in NLP. To the best of our knowledge, only a few small-scale studies in social sciences have attempted to automatically detect social groups. However, their results are not generalizable because they use only restricted varieties of text genres, and most of the datasets focus heavily on English (details in Section 2). Our work fills these gaps with three novel contributions. **(i)** We introduce the task of *group reference recognition*, and establish it as a token-level classification task. **(ii)** We release a new dataset GRIT (**G**roup **R**eference for **It**alian), the first Italian-language dataset for GRR. GRIT comprises a total of 169,566 tokens from multiple domains, where 22,855 tokens are manually identified as group ref-

[1]Earlier studies in the social sciences have already attempted to detect mentions of social groups (see Section 2). However, they did not focus on the task itself, but on the theoretical implications of group reference.

LREC-COLING 2024, pages 7963–7970

erence. The group references in GRIT cover a wide range of morpho-syntactic varieties, including 4,499 nouns, 1,807 proper nouns, and 223 relative clauses. **(iii)** We verify the potential of automatizing GRR with an experiment on GRIT using a fine-tuned BERT model. We believe that our contributions will facilitate future work on GRR and widen the landscape of NLP use cases.

2. Related Work

Linguistics Studies on Group References Within the domains of sociolinguistics and critical discourse analysis, considerable emphasis has been placed on examining how language usage influences identity formation across various social contexts (e.g., Eckert, 1989; Trudgill, 2000; Labov, 2006; Wodak, 2012; Fairclough, 2013). Further studies highlight the role of language in constructing social identities, notably through the establishment of "us" versus "them" dichotomy which effectively categorizes individuals and groups into insiders and outsiders (Wodak, 2014; Zotzmann and O'Regan, 2016). Within formal semantics, Barker (1992) explores the semantics of group terms such as *the committee* or *the group of women* in English, and argues that they denote atomic individuals rather than sets of members. Schwarzschild (1992) studies the semantics of plurals and their implications for understanding group references as well as their coreference possibilities. Schwarzchild argues that group references are not merely a collection of individuals, but entities *per se*, i.e., group references have specific characteristics that go beyond the sum of their individuals. Carlson and Pelletier (1995) extensively explore how generic expressions such as quantified noun phrases (e.g., *some students*, *every professor*) are interpreted and how they refer to classes or categories of entities. However, despite the acknowledged significance of social identities within discourse, none of these earlier studies directly focus on the automatic identification of social group references.

Social Science Studies on Group References Social scientists have found that the appeal to group identities plays a crucial role in inter-group political communication and social dynamics (Baker et al., 2008; Petrogiannis and Freidenvall, 2022). Within the framework of Social Identity Theory, Tajfel and Turner (2004) emphasize individuals' need to categorize themselves and others into distinct social groups, which consequently influences their behaviors and perceptions. Various studies on political communication have highlighted the role of group references in shaping policy appeals and political messages (Nteta and Schaffner, 2013; Horn et al., 2021; Howe et al., 2022). Indeed, political parties make sense to ordinary citizens only in terms of social identity (Achen and Bartels, 2017), and parties have been proven to extensively use social group references in their campaigns, with a focus on groups defined by ethnicity, social class, or gender (Huber, 2022). Further studies have confirmed that the attachment to a social group shapes how people think of themselves and how they vote (Bornschier et al., 2021; Hobolt et al., 2021).

Existing Studies on Automated GRR Even though there is clearly a substantial need for automatized GRR, only very few computational studies have attempted to tackle this task, and the language resources and tools they provide lack generalizability. Russmann (2020) rely solely on manual identification to identify group references. Haselmayer and Jenny (2017) and Decadri and Boussalis (2020) apply dictionary-based approaches, which have the drawback of not being able to detect out-of-vocabulary group references. In a more recent work, Licht and Sczepanski (2023) manually annotate group references in election manifestos of political parties in the UK and Germany, and fine-tune a RoBERTa model (Liu et al., 2019) to automatically detect them. However, their work is restricted to the genre of party manifestos, and focuses only on English and German. We extend this line of research by creating an Italian dataset that includes documents from diverse domains such as web-texts and parliamentary speeches.

3. Dataset Creation

3.1. Data Collection

For the creation of our dataset GRIT, we utilized raw texts from two sources: (i) the PAISÀ Corpus (Lyding et al., 2013), a corpus of Italian web-texts of diverse themes sourced from approximately 1,000 different websites; (ii) the Italian section of the corpus ParlaMint 2.1 (Erjavec et al., 2021), a multilingual corpus of parliamentary debates in Europe.

From both corpora, we randomly sampled documents that contain at least one term from a dictionary of 1,315 entries comprising the following categories of lexemes and their inflected forms:

(a) Job-related lexemes:
e.g., *idraulico* 'plumber', *insegnanti* 'teachers'

(b) Ethnicity-related lexemes:
e.g., *Francese* 'French', *Cristiani* 'Christians'

(c) Gender-related lexemes:
e.g., *donna* 'woman', *uomo* 'man'

We specifically selected these three categories because, following Huber (2022), sentences contain-

ing such lexemes are highly likely to also contain group references.

As the lengths of the documents in ParlaMint is on average around 10 times longer than those in PAISÀ (ParlaMint: 310.39 tokens on average; PAISÀ: 35.63 tokens on average), we sampled documents from PAISÀ and ParlaMint in a 9:1 ratio to achieve a balanced distribution between written texts and transcribed speeches. Furthermore, extracting documents from these two different sources strengthens the multi-domain and multi-genre nature of our new dataset.

3.2. Data Annotation

Before starting with the annotation process, we shuffled all documents to avoid any possible bias emerging from their order in the dataset. The initial annotation was conducted by two Italian native speakers ('annotators' henceforth), who are undergraduates of linguistics. The annotation was carried out using the software LightTag[2] (Perry, 2021). For each document, the annotators were asked to identify all the linguistic expressions that refer to individuals or groups of people, and mark those expressions with the label REF (Reference). For instance, the group reference 'the Sardinian senators' in (4) below should be labeled as the follows:

(4) *Maybe* [REF *the Sardinian senators*] *should have had the courage*.

In cases where the reference consists of complex syntactic structures such as relative clauses, the annotators were required to identify their minimal complete syntactic structure as exemplified in (5):

(5) *Those who work are tired.*
   a. Correct annotation:
   [REF *Those who work*] *are tired*.
   b. Wrong annotation:
   [REF *Those*] *who work are tired*.

As the primary aim of this work is to establish the task of GRR and to test the feasibility of automatizing it, in the current stage we annotated references in general (i.e., using only the label REF) and did not provide any fine-grained sub-categorization of group references.

The annotation involved multiple iterations: we divided all documents into 3 non-overlapping batches and annotated the batches sequentially. After each batch, we identified flaws in the annotation and refined the guidelines accordingly. The complete guideline is provided in Appendix A. The annotation process lasted 3 months for a total of 120 hours.

[2] https://www.lighttag.io/

3.3. Validation of Annotation

After the two annotators completed the initial annotation, a stringent review was conducted on the annotation results. Only the expressions that obtained unanimous agreement between the two annotators were immediately accepted without review. In cases where discrepancies emerged between the annotators (e.g., the annotators did not agree on the span of the expression, or only one annotator labeled a certain expression as REF), a review was conducted by a third Italian native speaker ('reviewer' henceforth), who is a graduate student of linguistics. This review process ensures a careful evaluation and resolution of the conflicts in the annotation results, contributing to the accuracy and consistency of the final annotation.

To evaluate the reliability of the annotations, we computed the overlap between every token annotated by (i) at least one of the two annotators and (ii) the reviewer. We achieved a Cohen's kappa of 0.82, indicating a very high level of agreement. Table 1 provides a summary of our dataset.

3.4. Dataset Accessibility

Our dataset GRIT is released as a free, public available dataset under the CC-BY-NC-SA 4.0 licence via GitHub: https://github.com/Sergio-E-Zanotto/grit

4. Experiment

To validate the feasibility of automatizing GRR, we conducted an experiment on GRIT using a fine-tuned BERT model (Devlin et al., 2019).

Data Selection and Preprocessing As group references are relatively infrequent (see Table 1), we pre-selected all sentences that contain at least one token labeled as REF. We also included sentences that do not contain any label REF, but contain at least one item from the dictionary we utilized to sample raw texts for dataset creation (see Section 3.1). In this way, we also incorporated representations of possibly ambiguous cases (e.g., the word 'French' in the dictionary can either have a group reference usage like in 'the French people', or a non-group-reference usage like in 'the French style').

From each span labeled as REF, we created two different versions: (i) a long version, which corresponds to the original span; (ii) a short version, which is automatically generated by removing the function words (mostly determiners) at the beginning of each span (e.g. *gli studenti* 'the students' is converted to *studenti* 'students'). The aim of experimenting with the short version was to investigate whether the model recognizes the importance of

Source	#Documents	#Sentences	#Tokens	#Tokens Labeled as REF
PAISÀ	900	3,074	85,623	14,128
ParlaMint	100	2,588	83,943	8,727
Total	1,000	5,662	169,566	22,855

Table 1: Overview of GRIT.

the initial function words in referential expressions. We were interested in this because (i) it is linguistically well established that determiners play a significant role in signaling referentiality (Carlson and Pelletier, 1995), and (ii) previous work has shown that Transformer-based language models tend to lack information about the semantics of function words (Kalouli et al., 2022).

Method We model the task of automated GRR as a binary token-level classification. All selected sentences were tokenized using *spaCy* (Honnibal and Montani, 2017), and for each labeled span, we mapped the label to every token. For instance, the sentence in (4) is mapped as follows, with `1` standing for tokens that are a part of a `REF`-span, and `0` vice versa:

(6)

Maybe	the	Sardinian	senators	should	[...]
0	1	1	1	0	0

We divided the selected data into training, validation and test sets in a 80/10/10 fashion (see Table 2). We fine-tuned the `bert-base-italian-cased` model[3] with a token classification head to achieve the classification (see Appendix B for setup details).

Results and Discussion Table 3 shows the results of our experiment. Even though the occurrence of group reference is relatively infrequent (percentage of group reference tokens in the test set: 17.60% for the long version and 12.09% for the short version, see Table 2), the fine-tuned BERT model can identify most of the instances. In both long and short versions, the accuracy is very high (0.96), whereas the precision and recall on the short version are slightly deteriorated by the absence of initial function words, reflecting the well-known role of determiners in referential expressions (Carlson and Pelletier, 1995). Overall, the results confirm the feasibility of automatizing GRR and its applicability on large-scale multi-domain Italian data.

Error Analysis Although the model achieves extremely good performance on the long sentences, in a qualitative error analysis we identified the following typical error types, which bring useful insights into the model's limitations in GRR: first, the model typically generates false positives on tokens that are ambiguous between group references and proper nouns of TV shows or locations. (7)–(8) below show such errors, where the false positive tokens are underlined. This can be attributed to the lack of world knowledge in the model.

(7) *Uomini e Donne*
'Men and Women' (an Italian TV show)

(8) *Santa Maria*
'Santa Maria (Maggiore)' (an Italian city)

Furthermore, the false negatives result typically from the model's poor performance in recognizing figurative group references, particularly those expressed through metonymy. (9)–(10) shows such cases, where 'minds' refer to people, and 'Austria' refers to the Austrian football team:

(9) [...] *menti ancora più chiuse*
'[...] minds which are even more closed'

(10) [...] *battendo in finale l'Austria*
'[...] beating in the final the Austria'

5. Conclusion and Future Work

GRR is a crucial task for research on political communication. Yet, its automatization has not gained much attention to date. So far only a limited amount of studies have attempted to automatize GRR, but they used non-generalizable dictionary approaches for detection or addressed only restricted genres. Addressing these research gaps, this paper aims to establish the task of GRR, and tests the feasibility of automatizing it and making it robust to unseen data. To this end, we introduced GRIT, the first multi-domain and multi-genre large-scale dataset for GRR in Italian. A token-level classification experiment on GRIT using a fine-tuned BERT model demonstrates GRR's great potential of being automatized and applied to social science studies with large datasets. Overall, the task of GRR extends the traditional study field of NER by tackling further complexities of referential linguistic expressions that NER does not cover.

In future work, we plan to enhance the annotation of group references by extending the current annotation scheme to a multi-label one with different sub-types, e.g., *ethnical group reference*, *gender group reference*, and *political group reference*.

[3] https://huggingface.co/dbmdz/bert-base-italian-cased

	#Sentences	#Tokens	#Tokens Labeled as REF (Long Version / Short Version)
Train	2,968	106,482	18,043 / 13,285
Validation	371	13,495	2,379 / 1,766
Test	372	13,821	2,433 / 1,792
Total	3,711	133,798	22,855 / 16,843

Table 2: Summary of the training, validation and test set (80, 10, 10).

	Accuracy	Precision	Recall	F1
Long	0.96	0.90	0.92	0.91
Short	0.96	0.86	0.90	0.88

Table 3: Classification results of fine-tuned BERT.

Moreover, we aim to tackle the more complex task of co-reference resolution for group reference, thus broadening the horizons of research on GRR.

6. Ethical Considerations and Limitations

Ethical Considerations The data collection and annotation are in line with the ethical regulations of the University of Konstanz (IRB 05/2021).

All annotators received a compensation of 15€/hour.

Limitations Given the complexity and novelty of the task GRR, the dataset GRIT primarily serves as a first attempt to test the feasibility of automatizing it. Thus, we only included a limited number of annotators. Furthermore, the experiment reported in Section 4, and the discussions of the model's limitations thereof, was based on only one model (BERT). Future work should consider comparing the performances and behaviors of different language models.

7. Acknowledgments

This project was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379. We thank Elisa Dervishi and Michele Bassanelli for their annotation work and their valuable insights into the refinement of the annotation process.

8. Bibliographical References

Christopher Achen and Larry Bartels. 2017. *Democracy for realists: Why elections do not produce responsive government*. Princeton University Press.

Paul Baker, Costas Gabrielatos, Majid Khosravinik, Michał Krzyżanowski, Tony McEnery, and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3):273–306.

Chris Barker. 1992. Group terms in English: Representing groups as atoms. *Journal of Semantics*, 9(1):69–93.

Simon Bornschier, Silja Häusermann, Delia Zollinger, and Céline Colombo. 2021. How "us" and "them" relates to voting behavior—social structure, social identities, and electoral choice. *Comparative Political Studies*, 54(12):2087–2122.

Greg N. Carlson and Francis Jeffry Pelletier, editors. 1995. *The Generic Book*. University of Chicago Press.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Silvia Decadri and Constantine Boussalis. 2020. Populism, party membership, and language complexity in the Italian chamber of deputies. *Journal of Elections, Public Opinion and Parties*, 30(4):484–503.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Penelope Eckert. 1989. *Jocks and Burnouts: Social categories and identity in the high school*. Teachers college press.

Norman Fairclough. 2013. Critical discourse analysis. In *The Routledge Handbook of Discourse Analysis*, pages 9–20. Routledge.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. Vancouver welcomes you! Minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1248–1259, Vancouver, Canada. Association for Computational Linguistics.

Martin Haselmayer and Marcelo Jenny. 2017. Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & quantity*, 51:2623–2646.

Sara B. Hobolt, Thomas J. Leeper, and James Tilley. 2021. Divided by the vote: Affective polarization in the wake of the Brexit referendum. *British Journal of Political Science*, 51(4):1476–1493.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Alexander Horn, Anthony Kevins, Carsten Jensen, and Kees Van Kersbergen. 2021. Political parties and social groups: New perspectives and data on group and policy appeals. *Party Politics*, 27(5):983–995.

Philip J. Howe, Edina Szöcsik, and Christina I. Zuber. 2022. Nationalism, class, and status: How nationalists use policy offers and group appeals to attract a new electorate. *Comparative Political Studies*, 55(5):832–868.

Lena Maria Huber. 2022. Beyond policy: the use of social group appeals in party communication. *Political Communication*, 39(3):293–310.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, coordination, and quantifiers in contextualized language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

William Labov. 2006. *The social stratification of English in New York city*. Cambridge University Press.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

Hauke Licht and Ronja Sczepanski. 2023. Who are they talking about? detecting mentions of social groups in political texts with supervised learning. OSF Preprints, 20 June 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tatishe Nteta and Brian Schaffner. 2013. Substance and symbolism: Race, ethnicity, and campaign appeals in the United States. *Political Communication*, 30(2):232–253.

Tal Perry. 2021. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vasileios Petrogiannis and Lenita Freidenvall. 2022. Political rhetoric, identities, and dominant gender representations: hegemonic masculinity in service of pro-austerity rhetoric in Greek political discourse. *Norma*, 17(2):88–106.

Uta Russmann. 2020. Voter targeting online in comparative perspectives: Political party websites in the 2008/2009 and 2013 Austrian and German election campaigns. *Journal of Political Marketing*, 19(3):177–200.

Roger Schwarzschild. 1992. Types of plural individuals. *Linguistics and Philosophy*, pages 641–675.

Kaare Strom. 1990. A behavioral theory of competitive political parties. *American Journal of Political Science*, pages 565–598.

Henri Tajfel and John C. Turner. 2004. The social identity theory of intergroup behavior. In *Political Psychology*, pages 276–293. Psychology Press.

Takehiro Teraoka. 2016. Metonymy analysis using associative relations between words. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4614–4620, Portorož, Slovenia. European Language Resources Association (ELRA).

Mads Thau. 2019. How political parties use group-based appeals: Evidence from britain 1964–2015. *Political Studies*, 67(1):63–82.

Peter Trudgill. 2000. *Sociolinguistics: An introduction to language and society*. Penguin UK.

Hao Wang, Siyuan Du, Xiangyu Zheng, and Lingyi Meng. 2023. An empirical study of incorporating syntactic constraints into BERT-based location metonymy resolution. *Natural Language Engineering*, 29(3):669–692.

Ruth Wodak. 2012. Language, power and identity. *Language Teaching*, 45(2):215–233.

Ruth Wodak. 2014. Critical discourse analysis. In *The Routledge Companion to English Studies*, pages 302–316. Routledge.

Karin Zotzmann and John P O'Regan. 2016. Critical discourse analysis and identity. In *The Routledge Handbook of Language and Identity*, pages 113–127. Routledge.

9. Language Resource References

Erjavec, Tomaž and Ogrodniczuk, Maciej and Osenova, Petya and Ljubešić, Nikola and Simov, Kiril and Grigorova, Vladislava and Rudolf, Michał and Pančur, Andrej and Kopp, Matyáš and Barkarson, Starkaður and Steingrímsson, Steinþór and van der Pol, Henk and Depoorter, Griet and de Does, Jesse and Jongejan, Bart and Haltrup Hansen, Dorte and Navarretta, Costanza and Calzada Pérez, María and de Macedo, Luciana D. and van Heusden, Ruben and Marx, Maarten and Çöltekin, Çağrı and Coole, Matthew and Agnoloni, Tommaso and Frontini, Francesca and Montemagni, Simonetta and Quochi, Valeria and Venturi, Giulia and Ruisi, Manuela and Marchetti, Carlo and Battistoni, Roberto and Sebők, Miklós and Ring, Orsolya and Dargis, Roberts and Utka, Andrius and Petkevičius, Mindaugas and Briedienė, Monika and Krilavičius, Tomas and Morkevičius, Vaidas and Diwersy, Sascha and Luxardo, Giancarlo and Rayson, Paul. 2021. *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*. CLARIN ERIC. PID http://hdl.handle.net/11356/1432.

Lyding, Verena and Stemle, Egon and Borghetti, Claudia and Brunello, Marco and Castagnoli, Sara and Dell'Orletta, Felice and Dittmann, Henrik and Lenci, Alessandro and Pirrelli, Vito. 2013. *PAISÀ Corpus of Italian Web Text*. Institute for Applied Linguistics, EURAC Research. PID http://hdl.handle.net/20.500.12124/3.

A. Annotation Guideline

Our complete annotation guideline is provided in Table 4.

B. Experimental Setup Details

The BERT model reported in Section 4 was fine-tuned for 2 epochs with a batch size of 16, a learning rate of 1e-4, and a weight decay of 1e-5. A random seed of 42 was used. The fine-tuning was implemented using the Hugging Face's *Transformers* library[4], and conducted on a NVIDIA A100 GPU with a total memory of 40GB. For the tokenization of the sentences, the model `it_core_news_sm`[5] from spaCy was used.

[4] https://huggingface.co/docs/transformers/index

[5] https://spacy.io/models/it

Annotation Guidelines: Group Reference Recognition in Italian

1. **Definitions**

   1.1 **Group reference**: all the tokens in the sentences that refer to people's identity (or to persons in general). E.g., *the teachers*, *a student*, etc.

   1.2 **Minimal meaningful syntactic phrase**: all the tokens that are fundamental to distinguish the group in the real world. E.g., *the son of Luigi of Denmark*.

2. **Task Description**
   In the documents, please mark all the tokens in the sentences that refer to either people's identity or persons in general. Please mark the entire minimal meaningful syntactic phrase that constitutes the reference.

3. **Guidelines and Examples**

   3.1 Label all the references found in the documents with 'REF'.

   3.2 Always mark the entire minimal meaningful syntactic phrase that constitutes the group reference. Examples:

   (1) a. ✓: *La figlia di Federico IV di Spagna*
   'The daughter of Federico IV of Spain'
   b. ✗: *La figlia di Federico IV di Spagna*
   'The daughter of Federico IV of Spain'

   3.3 If a group reference contains relative clauses, mark the entire relative clause together with its syntactic head. Examples:

   (2) a. ✓: *Io parlo a tutti quelli che non ce la fanno ad arrivare a fine mese*
   'I speak to all those who aren't able to make end meet at the end of the month'
   b. ✗: *Io parlo a tutti quelli che non ce la fanno ad arrivare a fine mese*
   'I speak to all those who aren't able to make end meet at the end of the month'

   3.4 Also mark group references expressed in figurative language, such as metaphors or metonymy. Examples:

   (3) a. Metaphor: *L'ondata di protesta si sposta verso la piazza*
   'The protesting wave is moving toward the square'
   b. Metonymy: *Il corpo fu rinvenuto la mattina*
   'The body was found in the morning'

   3.5 For proper nouns of organizations, mark them only when they refer to people composing the organization. Examples:

   (4) a. ✓: *Il Senato approva* 'The Senate approves'
   b. ✗: *La vicenda è avvenuta in Senato* 'The episode happened in the Senate'

   3.6 Do NOT mark references in titles of books, movies, competitions, etc. Example:

   (5) ✗: *Champions league*; *Il maestro e Margherita* 'The teacher and Margherita'

   3.7 Do NOT mark indefinite pronouns that do not have a clear reference. Examples:

   (6) a. ✓: *Nessuno dei votanti approva* 'Nobody of the voters approves'
   b. ✗: *Nessuno approva* 'Nobody approves'

Table 4: Annotation guideline.