Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07641-3

ORIGINAL ARTICLE

A multi-level methodology for the automated translation of a coreference resolution dataset: an application to the Italian language

Aniello Minutolo[1] · Raffaele Guarasci[1] · Emanuele Damiano[1] · Giuseppe De Pietro[1] · Hamido Fujita[2,3,4] · Massimo Esposito[1]

Received: 18 January 2022 / Accepted: 18 July 2022

Abstract
In the last decade, the demand for readily accessible corpora has touched all areas of natural language processing, including coreference resolution. However, it is one of the least considered sub-fields in recent developments. Moreover, almost all existing resources are only available for the English language. To overcome this lack, this work proposes a methodology to create a corpus for coreference resolution in Italian exploiting knowledge of annotated resources in other languages. Starting from OntonNotes, the methodology translates and refines English utterances to obtain utterances respecting Italian grammar, dealing with language-specific phenomena and preserving coreference and mentions. A quantitative and qualitative evaluation is performed to assess the well-formedness of generated utterances, considering readability, grammaticality, and acceptability indexes. The results have confirmed the effectiveness of the methodology in generating a good dataset for coreference resolution starting from an existing one. The goodness of the dataset is also assessed by training a coreference resolution model based on BERT language model, achieving the promising results. Even if the methodology has been tailored for English and Italian languages, it has a general basis easily extendable to other languages, adapting a small number of language-dependent rules to generalize most of the linguistic phenomena of the language under examination.

Keywords Coreference resolution · Corpus creation · Automated translation · Cross-language · Natural language processing · Linguistic phenomena

1 Introduction

Coreference resolution (henceforth CR) has a long history in natural language processing (NLP); knowing who is being talked about in a text has always been a fascinating challenge for scholars. Although it is not a new task, CR is still debated [1], demonstrating its usefulness concerning practical and theoretical issues. Indeed, coreference information has been used in various NLP tasks, such as text summarization [2], and also with reference to low-resource languages [3]. Moreover, it has been the object of study for linguistics theoretical issues [4], focusing on the interpretation of syntactic phenomena like null subjects and pronouns. Over the last decades, many approaches for CR have succeeded, ranging from simple rule-based systems to machine- and deep learning approaches [5, 6] to reinforcement learning-based solutions [7]. These approaches

✉ Raffaele Guarasci
raffaele.guarasci@cnr.it

[1] Institute for High Performance Computing and Networking of National Research Council of Italy (ICAR-CNR), Via Pietro Castellino 111, 80131 Naples, Italy

[2] Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam

[3] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain

[4] Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan

Published online: 19 September 2022

have also been transferred with applications to specific domains [8].

The history and developments in this field have led to the creation of numerous corpora specifically annotated for coreference-related tasks. From the earliest modestly sized corpora [9] manually created to progressively larger resources to satisfy the ever-increasing data needs of machine learning approaches [10] and capable of managing multiple languages or specific domains. Evaluation campaigns such as SemEval [11] and CoNLL 2012 [10] have contributed to the proliferation of available datasets. However, although it is long tradition in NLP, CR is one of the sub-fields of NLP, which has seen the slowest progress [1] during the last decade, dominated by the exponential growth of machine learning. In addition, the vast amount of resources available for the English language is not matched by a similar number for the other languages. Datasets in languages other than English are mainly limited to preexisting treebanks to which a specific coreference annotation level has been added.

About the language under investigation in this work, Italian, there are a few outdated annotated corpora [12–14], which suffer from limited size, excessive domain-dependence, and lack of a shared annotation standard scheme. Hence, only a handful of approaches for CR have been developed.

Starting from this issue, this paper describes an innovative cross-lingual methodology for creating a CR dataset in a low-resource language starting from a rich-resource one. The languages here considered are Italian and English, respectively. In particular, an Italian dataset for the CR has been generated starting from OntoNotes [15], which is currently considered the de facto standard for the evaluation of coreference tasks in English since the CoNLL shared tasks in 2011 and 2012.

The methodology is divided into two distinct steps. First, a multi-level translation process is applied to the English sentences extracted from the OntoNotes dataset for CR. This step aims to translate sentences trying to preserve mentions they can contain without losing in the translation the tokens composing the mentions, their positions, and the verbal agreements involving them. Second, a language refinement step has been introduced. This step tries to manage language-dependent phenomena to produce output sentences compliant with Italian grammar by applying language-specific rules derived from theoretical linguistics. These rules perform deletions and substitutions without losing information about mentions. Original coreference annotation has been preserved without having sentences that can sound unnatural or ungrammatical in Italian. This step is necessary in cases where there is a significant discrepancy between the two languages, in this case, Italian and English, concerning syntactic constructions involving personal pronouns that are often used in different ways.

Concerning evaluation, the results have been assessed both quantitatively and qualitatively. From the quantitative point of view, the readability of the produced sentences has been calculated using the Flesch–Kincaid index adapted for the Italian language [16]. This metric has been supplemented with a qualitative analysis carried out by native speakers using indicators from theoretical linguistics, such as grammaticality and acceptability. Grammaticality refers to a sentence's well-formedness from a syntactic point of view, e.g., the structure and order of the constituents are maintained. The concept of acceptability, instead, is related to how the sentence is considered semantically meaningful according to the annotator's judgments. Together, these two indicators allow assessing the quality of translated sentences from the perspectives of both grammatical correctness and meaningfulness for a native speaker. The goodness of the dataset has also been assessed by training a CR baseline model based on BERT [17]. Then, the results have been compared with the ones obtained by the same model but on the English version of the Ontonotes dataset.

The paper is organized as follows. Section 2 reviews the state of the art of datasets created for CR. It describes both the datasets created for English and other languages. Section 3 outlines the research motivations and contributions of the proposal. In Sect. 4, the methodology adopted for making the dataset starting from the original English resource is reported. This section describes the two macro-steps of translation and linguistic refinement to achieve a translated text that preserves mentions and coreferences. Section 5 discusses the results obtained, describing the evaluation process, both quantitative and qualitative, and outlining the performance achieved by a BERT-based CR model training on the generated dataset. Finally, Sect. 6 concludes the work.

2 Related work

The datasets developed over the years for CR are of various kinds. Generalist, domain-specific and multilingual datasets characterized by different criteria and annotation schemes have been created. The vast majority of the resources—as in all NLP fields—have been made for the English language, but there have also been developments in other languages in recent years.

It is worth noting that almost all resources include both coreference and anaphora resolution since both are part of the entity resolution family. The clear distinction in terminology between the two concepts is still debated in the literature. According to some studies, anaphora is a subset of coreference, while others claim that coreference is part

of anaphora. For this paper, the resources available for coreference will be listed, although these are almost always valid for anaphora resolution. Notice that—as regards terminology—in this work, the definition of coreference is the same as adopted in the OntoNotes schema. Therefore coreference is not limited to noun phrases [18] but includes pronouns, head of verb phrases and named entities as potential mentions.

Starting from these premises, this section first surveys and highlights the main characteristics of existing datasets for CR in English. Successively, CR resources for languages other than English are described, specifically outlining the ones for Italian.

2.1 CR resources for English

The MUC corpora is the first dataset manually created by human annotators that also aims for evaluation purposes. The MUC-6 [19] and MUC-7 [20] are based on North American news corpora (extracted by the Wall Street Journal), and they are small in size (318 annotated articles). Although now rarely used due to their limited domain and size, they are still considered valid compared to baselines. MUC has its evaluation metrics and SGML-based annotation format.

The GNOME Corpus [21] instead is created with a specific cross-domain scope. It includes texts from three domains (museum labels, pharmaceutical leaflets, and tutorial dialogues), and it has an annotation level of discourse and semantic information. GNOME has also been used in conjunction with other datasets to create the ARRAU corpus [22]. It includes corpora from different domains such as news-wire, dialogues, and fiction. The annotation scheme is the MMAX2 format which uses hierarchical XML files at the document and sentence level.

Then, there are corpora developed for specific coreference-related sub-tasks. The character identification corpus [23] focuses on the task of speaker-linking in multi-party conversations extracted from transcriptions of TV shows. ECB + [24] is another task-specific corpus. It is devoted to the topic-based event CR, a topic that has gained much attention in the literature in recent years.

Other corpora developed for cross-domain purposes exploit freely available online resources. The GUM corpus [25] is a multilayer, CoNLL-labeled corpus containing conversational, instructional, and news texts extracted from the web. WikiCoref [26] is composed of annotated Wikipedia articles, whose entities are linked to an external knowledge repository for the mentions. Both corpora use the OntoNotes schema for the annotation. It is worth noting that also the English Penn Treebank [27] has been used for purposes related to coreference tasks. Indeed, it was also annotated with coreference links as part of the OntoNotes project [15].

There are also coreference corpora specifically developed for a single domain. For instance, NP4E [28] is a small corpus based only on security and terrorism genres. It is annotated using the MMAX2 format for the event coreference task. In addition, the healthcare domain has received special attention, so numerous biomedical corpora have been created. Starting from GENIA corpus [29], which contains 2000 MEDLINE abstract, numerous other resources have been developed, such as Genia Treebank [30], Genia event annotation [31], and MedCo coreference annotation [32]. These resources have been the focus of the BioNLP-2011 shared task on Protein CR [33]. A different approach is proposed by CRAFT [34] and by its successor HANNAPIN corpus [35]. These resources contain full annotated biochemical articles for CR. In the pharmacological field, the DrugNerAR [36] corpus has been developed, with the aim of resolving anaphora for extraction drug–drug interactions in the pharmacological literature.

2.2 CR resources for other languages

The first corpus that also deals with languages other than English is ACE [37]. Initially based only on the journalistic domain, it aims to be heterogeneous and domain-independent and is annotated for different languages (like English, Chinese, and Arabic). The covered domains range from news-wire articles to conversational telephonic speech and broadcast conversations.

OntoNotes 5.0 [38] was the dataset involved in the Semeval 2010 [39] and CoNLL 2012 [10], with the aim of modeling CR for multiple languages. It was created to classify mentions of equivalence according to the entity to which they refer. OntoNotes is mostly based on news articles; it includes three different languages and is annotated using a CoNLL-like format. It is still the most widely used corpus for evaluation in the literature.

Another parallel corpus available in two languages (English and German) is ParCor [40]. It is a corpus that includes data extracted from a specific genre (TEDx talks and Bookshop publications). It focuses on a particular purpose: parallel pronoun CR in different languages in a machine translation context.

There are very few datasets currently used in the coreference task concerning the Italian language. VENEX [12] is a corpus which combines two different corpus-annotation initiatives: SI-TAL [41], focused on the creation of a corpus of written Italian from financial newspapers, and IPAR [42], which is a collection of spoken task-oriented dialogues of speakers. VENEX uses MATE as annotation scheme and MMAX for the markup.

Table 1 Size comparison of coreference corpora

Corpus	Language	Size (words) (k)
OntoNotes	English	1450
Venex	Italian	40
i-Cab	Italian	250
LiveMemories	Italian	250

Another coreference resource is I-CAB [13], a small dataset built on news documents taken from the regional newspaper *L'Adige*. Texts are annotated using a scheme derived from the ACE corpus.

The most recent corpus developed for Italian is LiveMemories [14]. It collects two genres of text: blog sites and Wikipedia pages related to the history, geography, and culture of the region of Trentino-Alto Adige/Südtirol. The annotation follows the ARRAU guidelines adapted for the Italian language. Table 1.

These resources present several limitations. First, they are related to a specific domain: both I-CAB and LiveMemories corpora contain only texts related to the region Trentin/Südtirol (respectively, newspaper articles and Wikipedia pages and blog sites). The VENEX corpus is more heterogeneous since it includes articles from financial newspapers and dialogues. Second, they adopt different annotation methods. VENEX annotation scheme implements the scheme proposed in MATE,[1] and the markup scheme is the simplified form of standoff adopted in the MMAX annotation tool. ICAB is annotated with a scheme inspired by the ACE corpus, while LiveMemories combines annotation methods from the ARRAU corpus for English [22] and the VENEX project.

3 Research objectives and contribution

The main objective of this work is to propose a cross-lingual methodology for the creation of a dataset for the CR by integrating an automatic translation and a rule-based refinement to transfer existing resources in a source language to a target language.

As highlighted in Sect. 2.1, the most recent datasets for coreference tasks are based on previously developed resources or treebanks to which an additional level of specific annotation has been added. This approach is practical for languages with a great richness of materials. Still, it cannot be adapted to languages like Italian, which are often overlooked in many NLP tasks due to limited resources.

Translating resources already developed in other rich-resource languages can address this shortcoming, but trying to maintain the same methodological accuracy used in creating the original dataset.

Translating existing datasets into other languages offers many advantages, considerably reducing creation time compared to creating a resource from scratch. This approach is not entirely straightforward. A fully automatic machine translation cannot be sufficiently accurate in adapting the original text to the linguistic features of the target language.

Therefore, as an element of novelty, the proposed methodology includes a step of language refinement derived from theoretical linguistics theory, particularly concerning aspects of syntax. This step tries to manage language-dependent phenomena to produce sentences compliant with the target language grammar and be perceived as correct by native speakers' judgements.

Despite this language-dependent refinement step, the proposed methodology has the character of reproducibility. It can be extended to other languages, developing a set of language-dependent refinement rules to generalize most of the linguistic phenomena of the language under examination. In addition, starting from existing resources makes it possible to obtain parallel corpora, useful for subsequent cross-lingual analysis.

From an application perspective, the proposed methodology has been used to create, to the best of our knowledge, the first medium-scale Italian dataset for CR that also respects properties of interoperability, domain independence, and compliance with annotation standards.

Indeed, the Italian language does not benefit from many resources, and, as highlighted in Sect. 2.2, existing material is outdated and restricted to VENEX [12], I-CAB [13], and LiveMemories corpora [14].

It is worth noting that both the excessive specificity of their application domains and their lack of a shared annotation standard scheme make interoperability between existing Italian resources extremely complicated. On the contrary, the corpus generated with the proposed methodology is comparable in size and annotation criteria with OntoNotes, which is currently considered the essential resource for the field [15]. The opportunity to compare with OntoNotes, which is the de facto standard for evaluating coreference tasks since the CoNLL shared tasks in 2011 and 2012, could open exciting perspectives for multilingual analysis.

The goodness of the generated dataset is also assessed concerning the possibility of being used to train a deep learning model for CR in Italian. To this aim, a baseline model on the dataset is generated by adopting a state-of-

[1] http://www.andreasmengel.de/pubs/mdag.pdf.

the-art deep learning architecture proposed for the same task in English.

4 Methodology for the creation of the dataset

The proposed cross-lingual methodology has been developed starting from the multilingual coreference annotation of the OntoNotes dataset first proposed by [10]. It is structured in two macro steps, as highlighted in Fig. 1. First, a coreference dataset is automatically translated from a source language into a target one, preserving mentions and their positions in texts. In detail, OntoNotes is used as an input coreference dataset expressed in English and Italian is selected as the target language.

A pipeline has been realized to perform this translation process.

In detail, first a CR dataset in the source language, denoted with α, is obtained from the *source corpus* by preserving documents, partitions, utterances, and mentions, but discharging irrelevant information and mentions whose tokens are contained in other mentions. Then, the dataset β_1 is obtained from the dataset α by discarding unwanted utterances, i.e., utterances lacking verbs or composed of too few or too many tokens. Successively, the dataset β_2 is obtained from the dataset β_1 by removing unwanted mentions, i.e., mentions that can easily lead to ambiguities and inaccuracies in their translation. After, the dataset β_3 is obtained from the dataset β_2 by removing all mentioned clusters within each partition resulting in inconsistency. Finally, the CR dataset γ in the target language is obtained from the dataset β_3 by translating its utterances and mentions through an intelligent token replacement/resolution procedure guided by the set $class(id_m)$, which is a set containing an estimation of the typology, gender and number of the real-world entities referred by each mention within the dataset β_3.

Second, a novel theoretical linguistics-based refinement is applied to improve the naturalness of the output text in the target language.

In particular, a series of rewriting rules based on principles of theoretical linguistics is applied to make it easy to obtain a more readable and fluent text in Italian from the original English text. The rules are structured in such a way as to ensure the most extensive coverage of the most frequent phenomena in the sentences. Subsequently, they have been automatically applied to the whole dataset.

Such rules are the most innovative aspect of the work of the methodology. Through the use of solid theoretical principles they allow enhancing the accuracy of a machine translation process on a specific task producing output sentences as close as possible to those produced by a native speaker in the target language.

In detail, first the dataset δ is obtained from the dataset γ by refining its utterances and mentions through a set of language-dependent refinement rules based on principles of theoretical linguistics to improve the naturalness and readability of the output text in the target language. Then, the final *output corpus* is obtained from the dataset δ by eventually rewriting pronouns and adjectives within utterances and mentions to improve their compliance to the target language concerning the agreement, inflexion, and subject–object role of grammatical constraints.

In the following, the characteristics of the input coreference dataset and two macro-steps of the methodology are diffusely explained.

4.1 Source corpus

The starting corpus in the source language is OntoNotes [15], a dataset containing primarily texts extracted from the news domain initially developed for the shared tasks on modeling unrestricted coreference at CoNLL 2011 [43] and CoNLL 2012 [10].

OntoNotes turns out to be an obligatory choice for many reasons. First of all, despite its lack of heterogeneity, it has a remarkable diffusion in the field, becoming the standard benchmark dataset used for CR. Even most recent systems perform the evaluation entirely on OntoNotes [44], although numerous other resources have been created for different domains.

OntoNotes also offers a considerable advantage in respect of size. As pointed out in Table 1 corpora currently available for the Italian language are pretty smaller. The size is a significant issue, primarily as it affects the possibility of using a corpus as the training set for a machine learning model.

Another reason lies in the annotation schema. As pointed out by several studies, one of the critical issues in corpus creation and annotation for the coreference task is the definition of the unit of text to be chosen as a mention of an entity.

This definition can depend on syntactic and semantic factors and involve several controversial problems discussed in theoretical linguistics. Coreference annotations of OntoNotes do not use the text (tokens) as a base layer, but they rely on a morpho-syntactic annotated layer. This feature relies on the fact that it is built on a hand-tagged treebank before the coreference dataset. The coreference portion of OntoNotes is not limited to noun phrases or a limited set of entity types. The aim of the project was to annotate linguistic coreference using the most literal interpretation of the text at a very high degree of

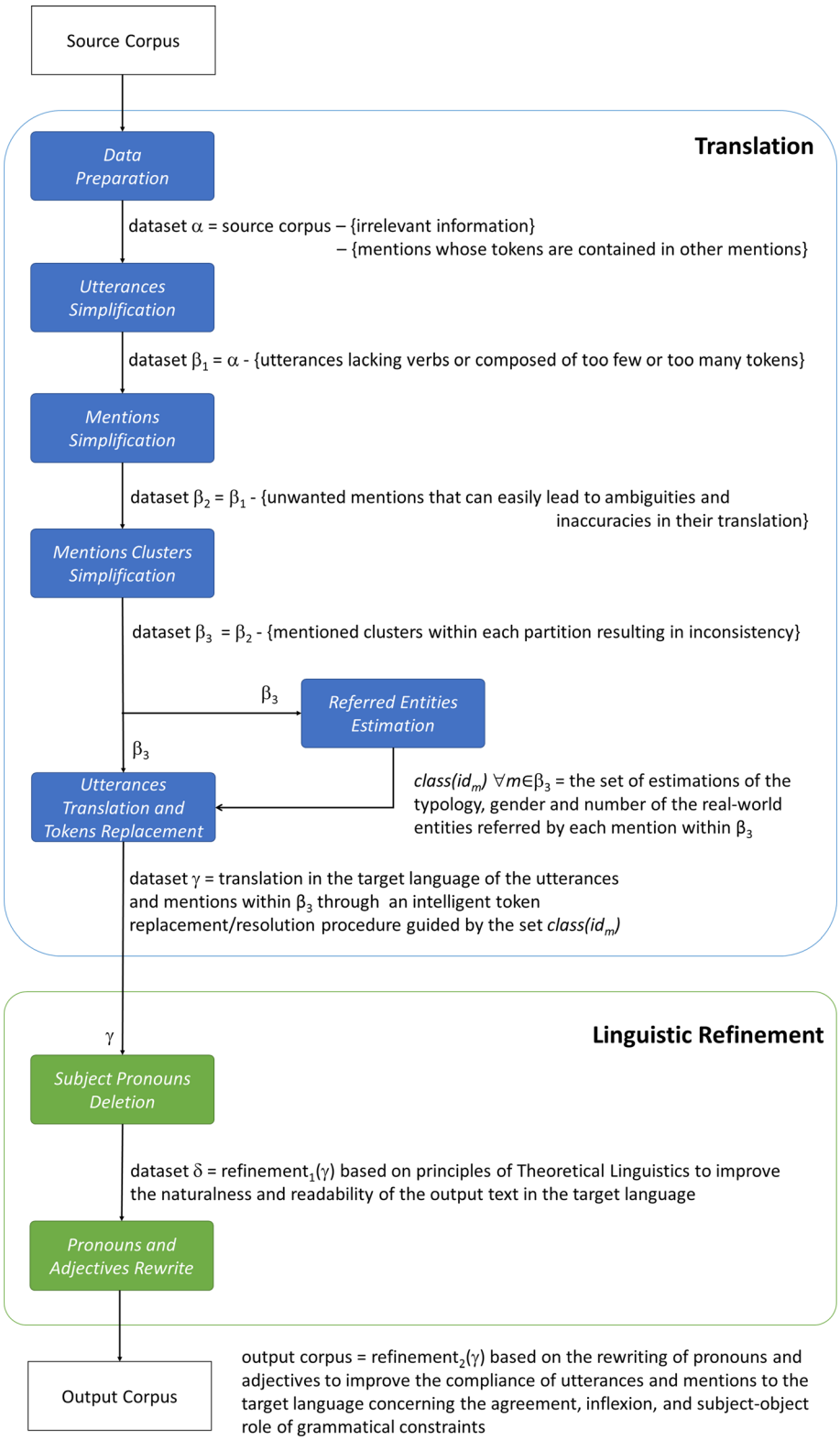

Fig. 1 The main steps of the proposed methodology

consistency, even if it meant departing from a particular linguistic theory [43].

The OntoNotes dataset is divided into three distinct subsets (*Train*, *Dev*, and *Test*), which can be used for training, developing, and testing a neural coreference model. The subsets *Train*, *Dev*, and *Test* are arranged into sets of documents composed of an ordered list of non-

Table 2 OntoNotes statistics

Measure	Train	Dev	Test
Total documents count	1940	222	222
Partitions for document	1.44	1.55	1.57
Maximum number of partitions in a document	23	21	28
Total partitions count	2802	343	348
Utterances for partition	26.83	28	27.24
Maximum number of utterances in a partition	188	127	140
Total utterances count	75,172	9603	9479
Utterances containing mentions	60,246	7420	7472
Maximum number of tokens in an utterance	210	186	151
Tokens for utterance	17.28	16.98	17.89
Mentions for utterance	2.07	1.99	2.09
Maximum number of mentions in an utterance	25	19	18
Coreference clusters for partition	12.54	13.25	13.024
Total coreference clusters count	35,143	4546	4532

overlapping partitions of ordered utterances. Statistics on the dataset are reported in Table 2.

Moreover, the distributions of the number of tokens and mentions per utterance in the OntoNotes dataset are reported in Fig. 2.

4.2 Translation

The translation step aims to extract, process, and correctly translate a dataset for CR, operating on both utterances and mentions contained in them.

As mentioned above, the input dataset is OntoNotes. It has been chosen as the best choice for this work. But it should be noted that any dataset for CR could also be utilized. The source and target languages have been English and Italian, even if almost all the considerations and procedures described in the following are valid or could be adapted to other languages.

In more detail, this step is first to extract from the dataset the set of linguistic information necessary for the translation. Second, the dataset is simplified by removing utterances, mentions, and mentions clusters not meeting some specific selection criteria. Third, unique replacement tokens are identified to be positioned in place of the mentions in the original utterances to preserve, after the translation, the tokens composing the mentions, their positioning, and the verbal agreements involving them. Lastly, the translation in the target language is performed. Mentions initially substituted by replacement tokens are also translated and reinserted in place of their corresponding translated replacement tokens, avoiding ambiguities due to more mentions made of the same token(s) in the same utterance.

In the following, more details are given about the whole translation process, breaking it down into six sub-steps, namely (1) data preparation, (2) utterances simplification, (3) mentions simplification, (4) mentions clusters simplification, (5) referred entities estimation, and (6) utterances translation and tokens replacement.

4.2.1 Data preparation

This step consists of a preliminary process to extract the information necessary to perform the following translation from the source dataset.

In detail, given **D** the set of documents in the source dataset, denoted with using $P(d) = [P_1,P_2,...,P_n]$ to denote the ordered list of non-overlapping partitions of utterances composing a document $d \in \mathbf{D}$, and denoted with $S(P) = [u_1,u_2,...,u_l]$ to denote the ordered list of utterances contained in a partition $P \in P(d)$, this step creates, for each utterance $u \in S(P)$, a quadruple $u' = (\mathbf{t}(u),\mathbf{p}(u),\mathbf{m}(u),s(u))$ where $\mathbf{t}(u)$ and $\mathbf{p}(u)$ are, respectively, the list of tokens composing u and their Penn Treebank POS (Part of Speech) tags, $\mathbf{m}(u)$ is the set of mentions built by selecting only the ones, eventually existing in u, containing no tokens of other mentions, and $s(u)$ is the label associated to the speaker of u.

An example of how the quadruple u' is built is reported in Fig. 3.

Only the mentions *"it"* and *"China"* are selected, whereas the mention *"an important city in China called Yichang"* is discharged since it contains tokens of a shorter mention, i.e., *"China."*

Each mention $m = (id_m,s_m,e_m)$ is a triple where id_m indicates the identifier of the referred real-world entity, s_m and e_m are the start and end indexes indicating the position of the tokens composing the mention in $t(u)$ and their POS tags in $p(u)$. Distinct mentions m_i and m_j are clustered when they refer to the same real-world entity, i.e., $id_{mi} = id_{mj}$, and

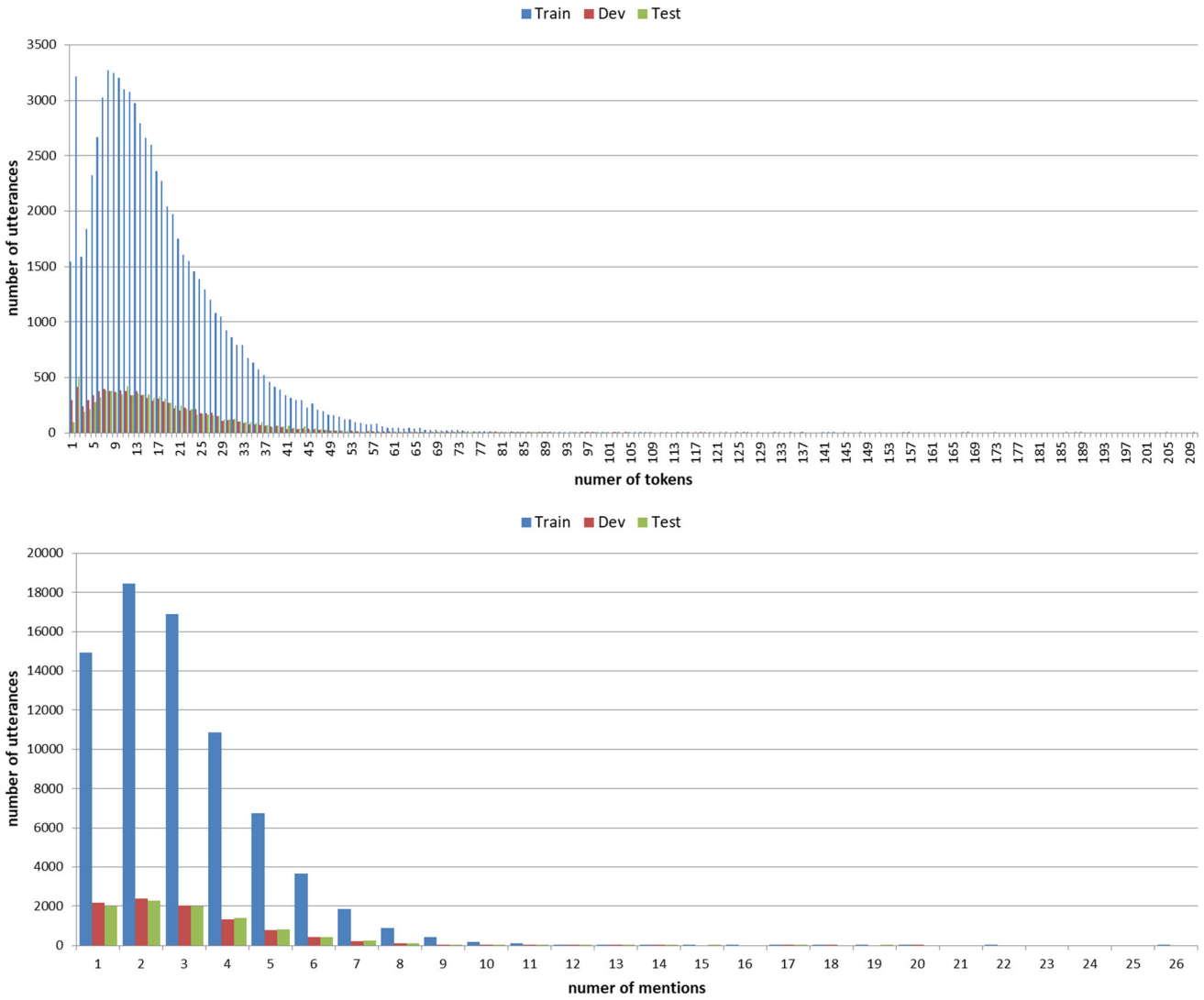

Fig. 2 Distributions of number of tokens and mentions per utterance in OntoNotes

only if they belong to the same document partition. More formally, given **I** the set of unique identifiers assigned to the real-world entities referred in a partition P of a document d, a cluster is defined as follows:

$$C(P \in \boldsymbol{P}(d), id \in \boldsymbol{I}) = \left\{ \bigcup_{u \in \boldsymbol{S}(P)} [(id_m, s_m, e_m) \in \boldsymbol{m}(u) : id_{m_i} = id] \right\}$$

Summarizing, starting from a source dataset containing n distinct documents d_1, d_2, ..., d_n, this step produces the following dataset α:

$$\alpha = \bigcup_{i=1}^{i=n} \left\{ \bigcup_{P \in \boldsymbol{P}(d_i)} \left[\bigcup_{u \in \boldsymbol{S}(P)} [(\boldsymbol{t}(u), \boldsymbol{p}(u), \boldsymbol{m}(u), s(u))] \right] \right\}$$

As an example, Fig. 4 reports a document partition within the dataset α and their associated mentions clusters.

It is worth noting that mentions belonging to different document partitions are assumed to refer to various real-world entities, i.e., the identifiers of real-word entities expire from one partition to another; thus, mentions clusters belonging to a partition are disjoint from the ones belonging to another partition.

4.2.2 Utterances filtering

This step is essentially devised to elaborate the dataset α to discard undesired utterances. In particular, first of all, given an utterance $u \in \alpha$, u is discarded or not in accordance with the criteria reported in Table 3.

This criterion derives from the consideration that, on the one hand, utterances containing no verbs or composed of a

few numbers of tokens should be discharged since they usually show missing or wrong grammatical dependencies. From a strictly linguistic point of view, Verbless sentences are more likely to be noun phrases instead of well-formed sentences. On the other hand, too long utterances often present a complex syntax resulting in difficult understanding even for a native-speaking human. The minimum and maximum thresholds used for selecting the utterances to be preserved have been chosen based on the Syntactic Capacity Limitation by human working memory and by computational language models for the correct understanding of the complex syntactic relations of a well-formed sentence [45].

A clarification on the terminology used is needed. For this work, the term *utterance* and *sentence* can be considered equivalent, although this is not precisely true in theoretical linguistics. OntoNotes only refers to utterances, which is why short sentences have been discarded in the proposed methodology. As mentioned above, short sentences tend to be not well-formed precisely because they are not technically sentences conveying a complete meaning. They are utterances, smaller units of speech which do not necessarily have a unit of meaning or a semantic structure.

Thus, the dataset β_1 is generated as follows:

$$\beta_1 = \alpha - \{u : u \in \alpha \wedge u \; is \; discharged\}$$

As an example, in Fig. 5 the same document partition shown in Fig. 4 is considered, where the utterance u_0 is discharged since it contains zero verbs. The utterances u_2, u_3 and u_5 are removed since they are composed of twenty-eight tokens resulting in intricate, not completely clear syntactic dependencies and hard to understand.

4.2.3 Mentions simplification

The dataset β_2 is generated by removing the undesired mentions from the dataset β_1, i.e., mentions that can easily lead to ambiguities and inaccuracies in their translation. To this end, both mentions composed of single or multiple tokens are evaluated by computing their dependency trees and using the roots to select the ones to be preserved on the basis only of the POS tags that can allow for estimating the gender and number of the referred real-word entities. (The estimation is performed in a subsequent step.)

It is worth noting that dependency tree roots coincide with mentions themselves if they are made of single tokens.

More formally, given a mention $m \in \beta_1$, denoted with $r_{t(m)}$ the root of the dependency tree of the tokens $t(m)$ composing m, m is discarded or not in accordance with the criteria reported in Table 4.

In particular, on the one hand, single-token mentions, as well as multi-token mentions containing zero verbs, whose dependency parse root, is a *Personal pronoun* in third person, a *Possessive pronoun* in third person, a *Determiner*, a *Noun*, or a *Proper noun*, are preserved. In contrast, in the other cases, they are discharged (note that, according to various studies [46] from 70 to 90% of the mentions are pronouns).

This choice is motivated by the fact that these kinds of mentions can enable the identification of gender and the number of real-world entities referred to by the mentions themselves, which, as a core idea of the proposed methodology, can support the preservation of the verbal agreements among the translated mentions and the other tokens within the translated utterance.

On the other hand, multi-token mentions containing one or more verbs are also discarded. Their dependency tree can be easily wrong, arising further ambiguities and

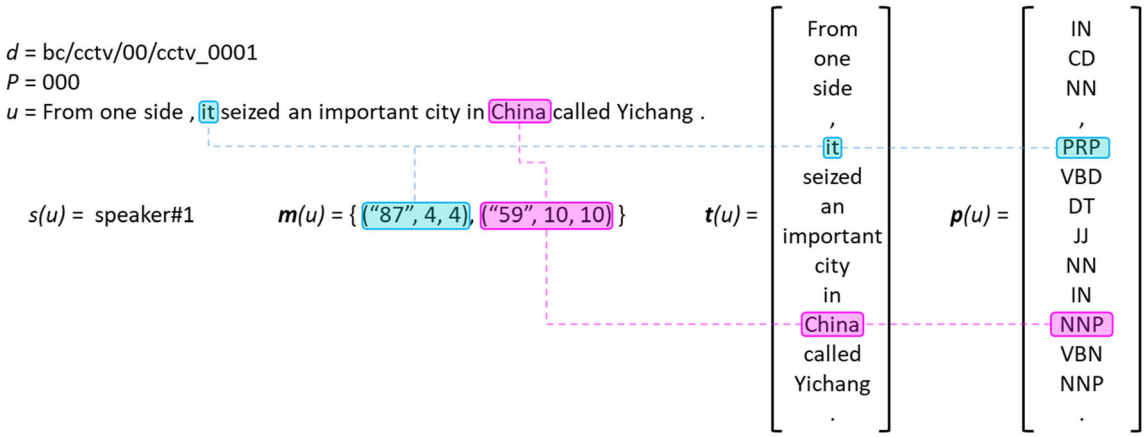

Fig. 3 Example of utterance contained in the dataset α

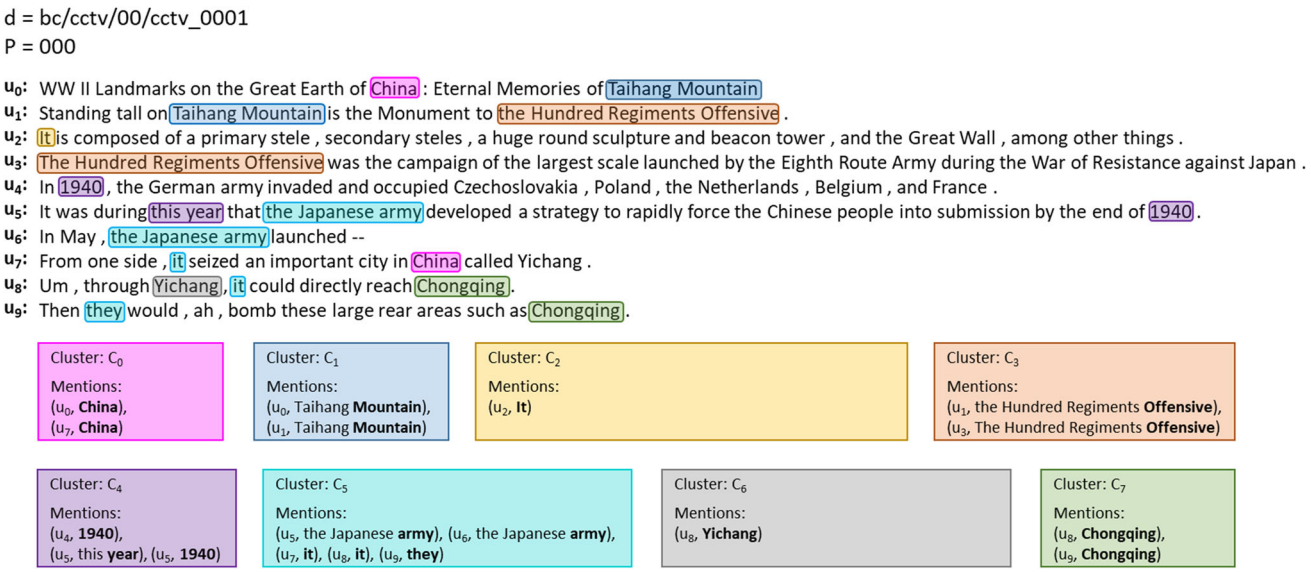

Fig. 4 Example of document partition and its mentions clusters

Table 3 The criteria followed for evaluating whether or not to preserve an utterance

IF u contains	and	THEN u is
1^+ verbs	$5 < \text{card}(u) < 21$	Preserved
1^+ verbs	Anything else	Discharged
0 verbs	Whatever	Discharged

inaccuracies in the process. Then, the dataset β_2 is generated as follows:

$$\beta_2 = \beta_1 - \{m = (id_m, s_m, e_m) : m \in \beta_1 \wedge m \; is \; discharged\}$$

An example related to the document partition shown above is reported in Fig. 6.

In particular, the mention *"China"* within the utterance u_0 is preserved since it is a *Proper noun*. On the contrary, *"1940"* is discharged since it is a *Numeral*. Moreover, the mentions *"Taihang Mountain"* and *"the Hundred Regiments Offensive"* within the utterance u_1 are preserved since their dependency trees exhibit as root, highlighted in bold, a *Proper noun*.

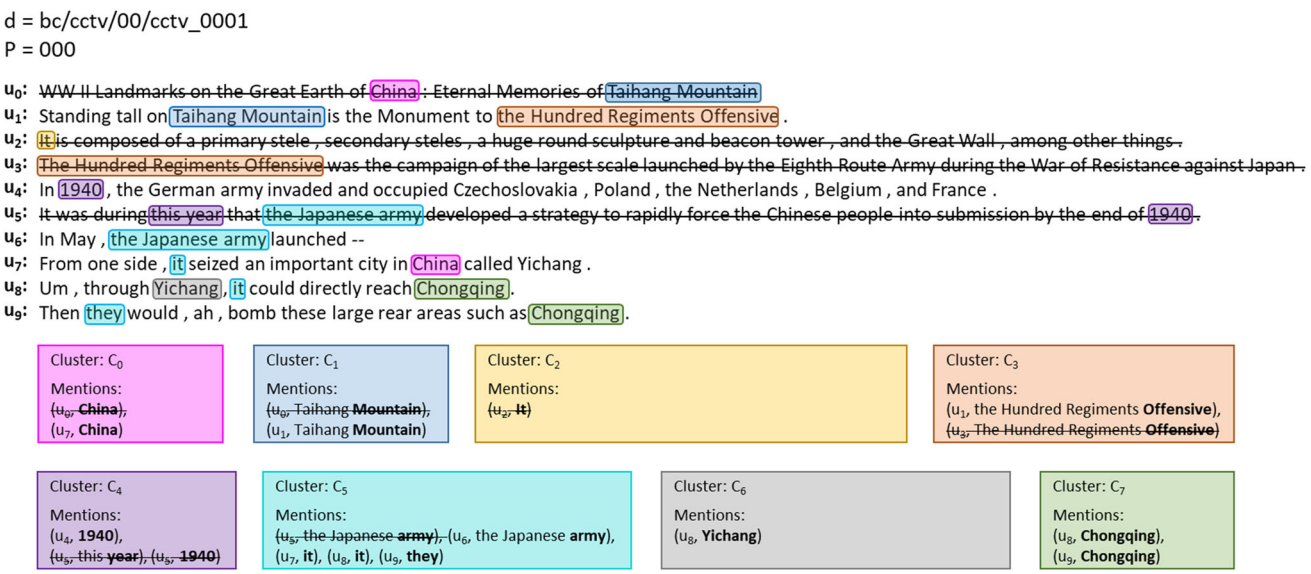

Fig. 5 Example of utterances from the dataset α not included in the dataset β_1

4.2.4 Mentions clusters simplification

The dataset β_3 is generated by removing from the dataset β_2 all mentioned clusters within each partition that are resulted in inconsistency after the previous utterances or mentions removals. More formally, a mentions cluster C is discarded or not according to the criteria shown in Table 5.

In detail, a mentions cluster C is preserved in the case when: (1) it is composed of at least two mentions; (2) it contains at least one mention whose dependency tree exhibits as root a *Noun* or a *Proper noun*. This second condition is meant to force the cluster to contain at least one mention capable of introducing a referred real-word entity. Clusters with zero elements after the previous removals are automatically discharged since they are meaningless. Then, the dataset β_3 is generated as follows:

$$\beta_3 = \beta_2 - \{m \in C : C \in \beta_2 \wedge C \text{ is discharged}\}$$

The same example document partition shown above is reported in Fig. 7, where some clusters are discharged.

In particular, the mentions cluster C_5 is preserved since it contains two elements, and one of them is the mention *"the Japanese army,"* whose dependency tree root is a *Noun*. On the contrary, the clusters C_0, C_1, C_3, and C_6 are discharged since their cardinality is less than two. For instance, the cluster C_0 becomes inconsistent after the previous removal of the utterance u_0 in the considered partition. The distribution of tokens and mentions per utterance in the dataset β_3 is reported in Fig. 8.

4.2.5 Referred entities estimation

This step aims to estimate the typology, gender and number of a real-world entity referred by a mention. This information will be used, in the next step, to determine unique replacement tokens to be positioned in place of the mentions to improve the overall translation by also preserving the verbal agreement.

More formally, given a mention $m = (id_m, s_m, e_m)$ within an utterance $u_s \in P$, with P is a document partition, this step is in charge of estimating the class $class(id_m)$ for each $m \in \beta_3$, where $class(id_m)$ is defined as the triple $(type(id_m), gender(id_m), number(id_m))$.

In detail, denoted with $t_t(m)$ the ordered list of tokens obtained after the translation of $t(m)$ in the target language, $class(id_m)$ is estimated by means of the following sequence of steps: (1) $r_{t(m)}$ is used to determine all the values for the triple $(type(id_m), gender(id_m), number(id_m))$; (2) in case some values for the triple cannot be determined from $r_{t(m)}$, $r_{tt(m)}$ in the target language is used; (3) finally, in case some values for the triple cannot be determined from both $r_{t(m)}$ and $r_{tt(m)}$, they are approximated referring to other mentions $m' \in \{C(P, id_m) - m\}$ belonging to the same cluster.

More precisely, in the case when $r_{t(m)}$ is a *Personal pronoun* or a *Possessive pronoun*, $class(id_m)$ is estimated as reported in Table 6:

In the last three rows, the gender of $class(id_m)$ cannot be determined immediately, and the other mentions belonging to the same cluster have been used to approximate it.

In case when the token $r_{t(m)}$ is a noun or a proper noun, the gender and the number of $class(id_m)$ cannot be directly deduced if the source language is English, since this information is not typically reported in the POS tags. Then, gender and number are derived from the POS tag generated for the token $r_{tt(m)}$ in the target language, if reported, or the other mentions belonging to the same cluster have used to approximate them.

Furthermore, in the case when the token $r_{t(m)}$ is a *Determiner*, the only way left is to approximate both gender and number referring to other mentions belonging to the same cluster.

The estimation of gender (number) for $class(id_m)$ from other mentions $m' \in \{C(P, id_m) - m\}$ belonging to the same cluster is performed by calculating the most frequent gender (number), giving more weight to the genders (numbers) suggested from pronouns than the ones

Table 4 The criteria followed for evaluating whether or not to preserve a mention m

IF m is	and $^{r}_{t}(m)$ is	THEN m is
Single token	A Personal pronoun in third person	Preserved
Single token	A Possessive pronoun in third person	Preserved
Single token	A Determiner	Preserved
Single token	A Noun or a Proper noun	Preserved
Single token	Anything else	Discharged
Multi token with 0 verbs	A Personal pronoun in third person	Preserved
Multi token with 0 verbs	A Possessive pronoun in third person	Preserved
Multi token with 0 verbs	A Determiner	Preserved
Multi token with 0 verbs	A Noun or a Proper noun	Preserved
Multi token with 0 verbs	Anything else	Discharged
Multi token with 1^{+} verbs	Whatever	Discharged

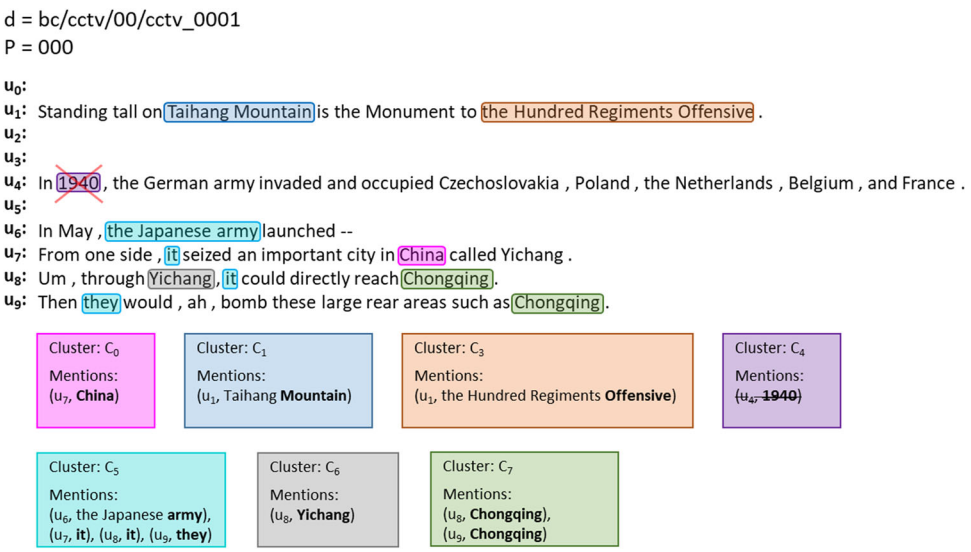

Fig. 6 Example of mentions from the dataset β_1 not included in the dataset β_2

Table 5 The criteria followed for evaluating whether or not to preserve a cluster C

IF	And	THEN m is
$card(C) \geq 2$	$\exists\ m \in C$: $r_{t(m)}$ is *Proper noun* or *Noun*	Preserved
$card(C) \leq 1$	Whatever	Discharged

suggested from the nouns. More formally, gender and number of $class(id_m)$ are determined as reported in Table 7 and in Table 8, respectively.

As an example, consider two utterances u_1 = *"Lora Owens is the stepmother of Albert Owens ."* and u_2 = *"She joins us now by phone ."* belonging to the same document partition, and the mentions cluster $C_6 = \{m_1 \in u_1, m_2 \in u_2\}$, where m_1 = *"Lora Owens"* and m_2 = *"She."* The $class(6_{m2})$ can be easily determined as equal to *(human, female, singular)*, whereas, on the contrary, no information can be inferred for $class(6_{m1})$ by evaluating the mention m_1. Thus, it can be estimated on the basis of the values of the other mention m_2 belonging to C_6. Roughly speaking, since the cluster C_6 contains one pronoun suggesting that the referred real-word entity is a female human, then this information can be extended also to the other mention to estimate its class.

4.2.6 Utterances translation and tokens replacement

This step is devised to perform a sequence of three actions on each utterance $u_s \in \beta_3$ expressed in the source language,

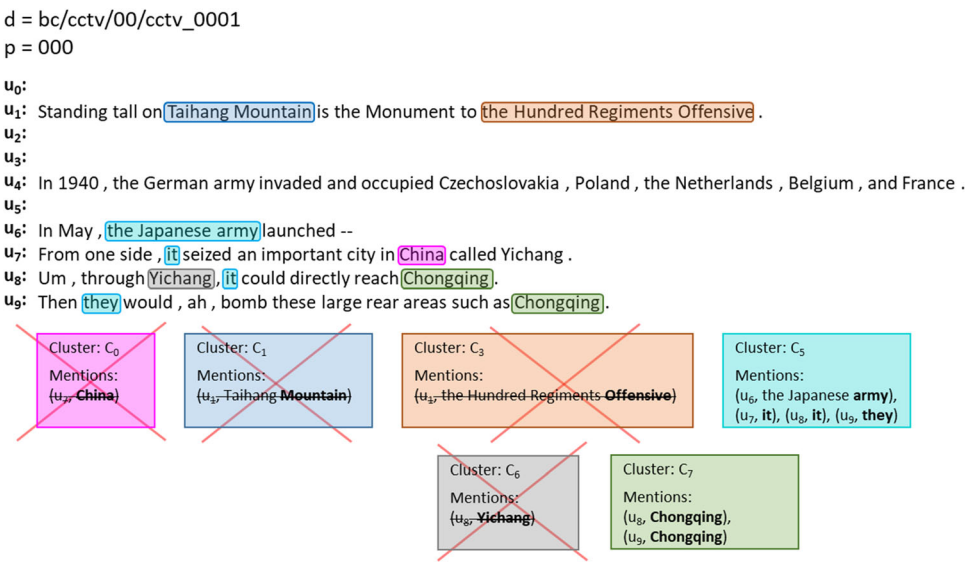

Fig. 7 Example of clusters from the dataset β_2 not included in the dataset β_3

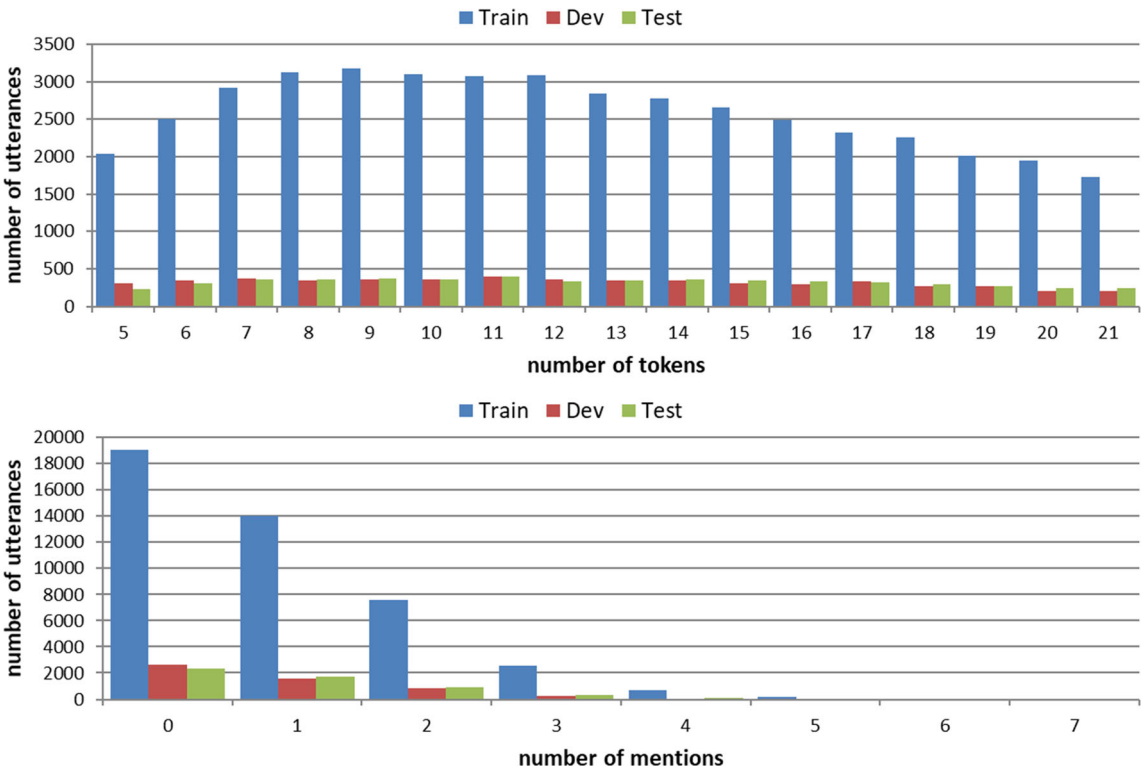

Fig. 8 Distribution of tokens and mentions per utterance in the dataset β_3

namely tokens replacement, utterances translation and tokens resolution.

First, tokens replacement consists in of evaluating, for each mention $m \in u_s$, the triple $class(id_m)$, in order to select a unique token $m' \in / u_s$ to be positioned in place of m and, as a result, generate the utterance u'_s. It is worth noting that, $u'_s = u_s$ in the case when no mention is contained in u_s.

Replacement tokens are randomly extracted from predefined lists of unique tokens built such that, on the one hand, they exhibit the same type, gender, and number of the $class(id_m)$ and, on the other hand, their representations in the source and target language are the same, i.e., $r_{t(m')} = r_{tt(m')}$. This choice increases the chance that replacement tokens appear unchanged within a translated utterance.

As an example, the utterance u_s = *"Lora Owens is the stepmother of Mary White, she joins us now by phone."* contains three mentions m_1 = *"Lora Owens,"* m_2 = *"Mary White,"* and m_3 = *"she."* In the hypothesis that $class(m_1) = class(m_2) = class(m_3)$ = *(human, female, singular)*, three replacement tokens m'_1 = *"Gabriella,"* m'_2 = *"Serena,"* and m'_3 = *"Sabrina"* are selected from a list of women's names whose representations in the source and target language are the same. These tokens are positioned in place of m_1, m_2, and m_3 and, as a result, the utterance u'_s = *"Gabriella is the stepmother of Serena, Sabrina joins us now by phone."* is generated.

Second, utterances translation consists, on the one hand, in generating the utterance u'_t by translating u'_s in the target

Table 6 The estimation of $class(id_m)$ for Personal and Possessive pronouns

IF $r_{t(m)}$ is equal to	THEN $class(id_m)$ is
"she", "her", "hers", or "herself"	(human, female, singular)
"he", "him", "his", or "himself"	(human, male, singular)
"it", "its", or "itself"	(thing, ?, singular)
"they", "their", or "theirs"	(thing, ?, plural)
"them" or "themselves"	(thing, ?, plural)

Table 7 The estimation of the gender of $class(id_m)$ for a mentions cluster

IF within the cluster	THEN $gender(id_m)$ is
Female pronouns $\geq$ male pronouns	Female
Male pronouns $\geq$ female pronouns	Male
Female nouns $\geq$ male nouns	Female
Male nouns $\geq$ female nouns	Male
Otherwise	Male

Default setting is "male" due by occurrences in the corpus

Table 8 The estimation of the number of $class(id_m)$ for a mentions cluster

IF within the cluster	THEN $number(id_m)$ is
Singular pronouns $\geq$ plural pronouns	Singular
Plural pronouns $\geq$ singular pronouns	Plural
Singular nouns $\geq$ plural nouns	Singular
Plural nouns $\geq$ singular nouns	Plural
Otherwise	Singular

language and, on the other hand, verifying for each $m' \in u'_s$ its existence in u'_t. In case when $\exists m' \in u'_s : m' \notin u'_t$ the token replacement is performed again for the utterance u_s and a distinct token for m is selected.

As an example, for the utterance u'_s = *"Gabriella is the stepmother of Serena, Sabrina joins us now by phone."*, the utterance u'_t = *"Gabriella 'e la matrigna di Serena, Sabrina si unisce a noi ora per telefono."* is generated.

Third, tokens resolution consists in of generating, for each mention $m \in u_s$, the tokens $r_{tt(m)}$ by translating $r_{t(m)}$ in the target language. Moreover, the utterance u_t is also generated from u'_t by resolving each m' within u'_t through the positioning of the tokens $r_{tt(m)}$ in place of m'. It is worth noting that $u_t = u'_t$ in the case when no replacement token is contained in u'_t.

As an example, given the utterance u'_t = *"Gabriella 'e la matrigna di Serena, Sabrina si unisce a noi ora per telefono."* the replacement tokens m'_1 = *"Gabriella,"* m'_2 = *"Serena,"* and m'_3 = *"Sabrina"* are resolved on the basis of the translated tokens $r_{tt(m1)}$ = *"Lora Owens,"* $r_{tt(m2)}$ = *"Maria Bianca"* and $r_{tt(m3)}$ = *"lei"* and, as a result, the utterance u_t = *"Lora Owens 'e la matrigna di Maria Bianca, lei si unisce a noi ora per telefono."* is generated.

As a result of this step, the dataset γ is generated.

4.3 Linguistic refinement

This step is in charge of applying to the dataset γ a set of language-dependent refinement rules based on principles of theoretical linguistics to improve the naturalness and readability of the output text in the target language.

It is necessary to make some minor clarifications about the differences between the two languages under analysis from a linguistic point of view. Italian and English have multiple differences, beginning with their origin, the variability of constituents in word order, and greater or lesser morphological richness. First of all, English is a Germanic language with rigid word order, and extremely small inflectional variation [47], its fixed subject–verb-order structure implies a mandatory explicit subject. By contrast, Italian belongs to the Romance subgroup of Italic languages, characterized by high verbal inflection and [48] great freedom in the order of constituents [49]. Such morphological richness leads to a different configuration of syntactic structures involving pronouns. In particular, it results in the omission of the subject pronoun. As pointed out by recent studies, this misalignment produces difficulties in the translation process since the missing pronoun is challenging to be reproduced, and it affects the order of dependencies in the sentence [50, 51]. For that reason, from a practical point of view, the refinement rules for the target language have been focused on improving the use of personal and possessive pronouns and, in addition, of possessive and demonstrative adjectives.

Indeed, generally speaking, personal and possessive pronouns often represent the primary part of the discourse used to co-refer to an entity, as reported in [46, 52]. Moreover, also for the dataset γ, a greater distribution of pronouns as single-token coreferences is observed and confirmed, as reported in Table 9.

In more detail, in Italian, two specific phenomena typically occur altering the use of pronouns and adjectives concerning English, namely *null-subject* and *agreement and morphemes inflexion*.

Null-subject phenomenon permits an independent utterance to lack (or lack) an explicit subject. Such truncated utterances have an implied or suppressed subject that can be determined from the context. In particular, null subject languages, like Italian, express person, number, and/or gender agreement with the verb inflexion, making a subject noun phrase redundant. It is worth noting that the lack of an explicit subject does not create an utterance ungrammatical, but it is often perceived as less natural by native speakers. As an example, in the utterance *"Giovanni and'o a far visita a degli amici. Per la strada, [egli] compr'o del vino"* (*"Jonh went to visit some friends. On the way, [he] bought some wine"*) the subject pronoun *"egli"* (*"he"*) is suppressed in Italian. This phenomenon is not present in English and the strategy of coreference annotation used in OntoNotes for the pronouns is difficult to completely match with a language belonging to a different linguistic family, such as Italian, where pronouns can be omitted when used as the subject in an utterance. Notice that the translation involving *null-subject* languages is still a heavily debated issue in the literature because of the difficulty in representing dropped pronouns [51]. In recent years many studies have addressed the problem proposing different solutions for different languages [53–55], including Italian [56].

On the other hand, *agreement* is a morpho-syntactic phenomenon in which the gender and number of the subject and/or objects of a verb must also be indicated by the

verbal inflexion. As an example, consider the utterance *"Quello 'e andato"* (*That one is gone*), where the singular masculine pronoun subject *"quello"* (*that one*) is agreed with the past participle *"andato"* (*gone*) of the verb *"andare"* (*go*) to which it refers. The past participle, indeed, presents a singular masculine inflexion, as highlighted by the suffix *-o*. Therefore, the correct suffix of the pronoun is the same as that of the noun *"quello."*

In English, pronouns and adjectives do not exhibit any inflection, and thus, their agreement with the verbs is not expressed. On the contrary, they must be in concordance with the verbal forms in Italian. As a result, after translating both pronouns and adjectives from English to Italian, their agreement with the verbs must be verified and granted if it is not respected.

In summarizing, this step of linguistic refinement is meant to further refine the dataset γ by removing not mandatory subject pronouns and rewriting pronouns and adjectives to grant correct agreement and inflexions. In the following, more details are given, breaking this step down into two sub-steps, namely (1) Subject Pronouns Deletion and (2) Pronouns and Adjectives Rewrite.

As a result of this step, the dataset δ is generated.

4.3.1 Subject pronouns deletion

This step aims to properly handle the null subject phenomenon for the pronouns occurring in dataset γ after the translation in the target language, i.e., Italian. It is in charge of evaluating the utterances within γ to (1) delete personal pronouns assuming the subject role in them; (2) move any mention associated with deleted subject pronouns on the verbs in dependency relation with them.

More formally, given an utterance $u \in \gamma$, denoted with $DT(u)$ its dependency tree, with t_i and t_j the ith and jth elements of the list of tokens $t(u)$, with $d(t_j,t_i) \in DT(u)$ a dependency relation from t_j to t_i, and with $label(d)$ the label associated with the typed dependency relation d, the criteria followed for performing the pronouns deletion are reported in Table 10.

In detail, first, a personal pronoun is identified as the subject of a clause contained in an utterance $u \in \gamma$ by verifying if it is connected with a verb through a direct grammatical dependency, typed as subject, in the corresponding dependency tree. Each personal pronoun labeled as subject can be removed.

If no mention is placed on the subject pronoun to be deleted, it is simply removed from the utterance. On the contrary, in case a mention is positioned on it, the mention is moved toward the verbal constituent it is dependent on, as calculated in the corresponding dependency tree, following the approach proposed in MATE Guidelines [57] and LiveMemories Corpus [14].

As an example, in the utterance *"[Egli] ha detto alla gente che [lei] era una brava cuoca"* (*"[He] has told people that [she] was a good cook"*), the personal pronouns *"Egli"* and *"lei"* act as subjects of their clauses and can be omitted. The deletion of the subject pronouns *"Egli"* and *"lei"* generates the shift of the mentions placed on them toward the verbal constituents *"ha"* (*"has"*) and *"era"* (*"was"*) on which they are dependent.

Table 9 Distribution of most frequent single-token coreference POS in the dataset γ

Part-of-speech	Percentage
Pronouns	33.6
Proper nouns	10.6
Nouns	8.6
Determiners	6.8
Verb	1.08
Adverbs	1.04
Adjectives	0.8

4.3.2 Pronouns and adjectives rewrite

This step aims to evaluate each utterance to identify pronouns and adjectives that can be rewritten to improve their compliance to the Italian language concerning the agreement, inflexion, and subject-object role of grammatical constraints.

The first set of rules operate on personal pronouns in clauses verifying and correcting (1) their agreement in number with verbs, in case they assume the role of subjects, and (2) the correspondence between the syntactic role (subject or object) and the inflected form (first or second singular person). More formally, given an utterance $u \in \gamma$, denoted with textitnumber($t \in t(u)$) the number of a token t indicating if t is expressed, or is assigned to be, in its singular or plural form, the criteria adopted to rewrite personal pronouns are reported in Table 11.

In detail, in the first rule, a personal pronoun t_i is identified as the subject of a clause contained in an utterance $u \in \gamma$ by verifying if it is connected with a verb t_j through a direct grammatical dependency $d(t_i,t_j)$, typed as subject, in the corresponding dependency tree. Then, the agreement in number between the subject pronoun and the corresponding verb is verified and possibly corrected. As an example, the utterance *"Tu siete nella stanza"* (*"You are in the room"*) contains the personal pronoun in second person singular *"Tu"* (*"You"*) in disagreement with the plural form of the verb *"siete"* (*"are"*). Thus, the personal pronoun is rewritten in the plural form as *"Voi."*

In the next two rules, personal pronouns in first or second singular person are verified if preceded by a preposition and wrongly assuming the form of the subject pronoun, i.e., *"io"* (*"I"*) and *"tu"* (*"you"*), and corrected

Table 10 The criteria followed for performing pronouns deletion

IF	and	and	and	THEN
$t_i \in t(u)$ is Personal pronoun	$t_j \in t(u)$ is aux r verb	$\exists d(t_j,t_i) \in DT(u)$: $label(d) = subject$	$\exists m \in m(u)$: $s_m = e_m = i$	$s_m = e_m = j$ $t(u) = t(u) - t_i$
$t_i \in t(u)$ is Personal pronoun	$t_j \in t(u)$ is aux or verb	$\exists d(t_j,t_i) \in DT(u)$: $label(d) = subject$	otherwise	$t(u) = t(u) - t_i$

Table 11 The criteria followed for rewriting personal pronouns

IF $t_i \in t(u)$ is	and $t_j \in t(u)$ is	and	THEN
Personal pronoun	aux or verb	$\exists d(t_j,t_i) \in DT(u)$: $label(d) = subject$	$number(t_i)$ is $number(t_j)$
Personal pronoun *"io/tu"*	Preposition	$j = i - 1$	t_i is *"me/te"*
Personal pronoun *"me/te"*	conjunction *"che"*	$j = i - 1$	t_i is *"io/tu"*

with the corresponding form for the object role, i.e., *"me"* (*"me"*) and *"te"* (*"you"*).

In the last two rules, personal pronouns in first or second singular person are verified if preceded by the conjunction *"che"* (*"that"*) and wrongly presenting their object pronoun formho *"me"* (*"me"*) and *"te"* (*"you"*), and corrected with the corresponding subject role, i.e., *"io"* (*"I"*) and *"tu"* (*"you"*). As an example, the utterance *"Non credono che me sia pronto"* (*"They do not think me am ready"*) wrongly uses the pronoun *"me"* in its object role. Thus, it is rewritten as *"io"* (*"I"*) since it has a subject role in the clause introduced by the conjunction *"che"* (*"that"*).

The second set of rewrite rules evaluates the agreement in gender and number between possessive and demonstrative adjectives and the noun they refer to (typically the noun before or immediately after them), following the criteria reported in Table 12.

In detail, in the first rule, each possessive adjective t_i within an utterance $u \in \gamma$, is identified as connected with a noun t_j by means a direct grammatical dependency $d(t_i,t_j)$, typed as possessive determiner, in the corresponding dependency tree. Then, the agreement in gender and number between the possessive adjective and the corresponding noun is verified and possibly corrected. As an example, the utterance *"Mia padre lavora in banca"* (*"My father works in a bank"*) contains the possessive adjective *"Mia"* (*"My"*) with the feminine suffix *"-a"* in disagreement with the male singular noun *"padre"* (*"father"*), (while the corresponding *"my"* in English has no inflection). Thus, it is rewritten as *"mio"* with masculine suffix *"-o."*

In the second rule, each demonstrative adjective t_i within an utterance $u \in \gamma$ is recognized as related to a noun t_j if this latter occurs at most four tokens forward and is connected through a direct grammatical dependency $d(t_i,t_j)$, typed as a generic determiner, in the corresponding dependency tree.

Then, the agreement in gender and number between the demonstrative adjective and the corresponding noun is verified and possibly corrected. Moreover, the suffix of the demonstrative adjective t_i is also checked and modified on the basis of the initial letters of the token $t_i + 1$ immediately following t_i in u, as reported in Table 13.

The thresholds concerning minimum and maximum tokens number and the distance of demonstratives are inspired by recent studies [58] that have quantitatively estimated the syntactic capacity limitation by human working memory and by computational language models for the correct understanding of the complex syntactic relations of a well-formed sentence.

As an example, the utterance *"Quella avviso è stato redatto nelle ultime 24 ore ."* (*"That notice has been drafted in the last 24 hours ."*) contains the demonstrative adjective *"Quella"* (*"That"*) with feminine suffix *"-a"* in disagreement with the male singular noun *"avviso"* (*"avviso"*). Thus, the demonstrative adjective is rewritten as *"Quello"* with masculine suffix *"-o."* Moreover, since the token following the demonstrative starts with a vowel, *"Quello"* is further replaced with its elided form (*"Quell'"*).

Finally, the last typology of rewriting rule evaluates if a demonstrative is used as a pronoun and replaces it with a neuter term following the criteria reported in Table 14.

In detail, this rule evaluates if a demonstrative t_i within an utterance $u \in \gamma$, is related to a noun t_j in a span of maximum 4 tokens. In the negative case, it is assumed to work as a pronoun and, thus, it can be replaced by a neuter term preventing possible agreement errors in long-distance syntactic dependencies. As an example, the utterance *"Quella è stato fatto nelle ultime 24 ore ."* (*"That has been done in the last 24 hours ."*), contains the

Table 12 The criteria followed for rewriting possessive and demonstrative adjectives

IF $t_i \in t(u)$ is	and $t_j \in t(u)$ is	and	THEN
Possessive	Noun	$\exists d(t_j, t_i) \in \boldsymbol{DT}(u) : label(d) = possessive$	$gender(t_i)$ is $gender(t_j)$
Adjective		*determiner*	$number(t_i)$ is $number(t_j)$
Demonstrative	Noun	$\exists d(t_j,t_i) \in DT(u)$:	$gender(t_i)$ is $gender(t_j)$
Adjective		$label(d) = determiner \wedge$	$number(t_i)$ is $number(t_j)$
		$i < j \leq i + 4$	suffix(t_i) is set based on t_z[0] and t_z [1] where $z = i + 1$

Table 13 The criteria followed for rewriting possessive and demonstrative adjectives

Gender(t_j) is	Number(t_j) is	First letter of t_{i+1} is	Modified form of t_i is
Masculine	Singular/Plural	Any voxel	Quell'/Quegli
	Singular/Plural	S + consonant, PS, GN, X, Y, Z	Quello/Quegli
	Singular/Plural	Other consonants	Quel/Quei
Feminine	Singular/Plural	Any voxel	Quella/Quelle
	Singular/Plural	Any consonant	Quell'/Quelle

demonstrative *"Quella"*(*"That"*) which is not connected with a noun in a span of maximum four tokens. Thus, it is replaced by the neuter demonstrative *"Ciò"*(*"That"*).

Notice that all rules aimed at rewriting, deleting, and modifying mentions do not affect the complexity of the language under consideration. The rules do not simplify syntactic phenomena or grammar, but they try to respect the syntax of the target language (Italian) without losing the information on mentions and co-references present in the source language (English).

5 Results and evaluation

The dataset δ obtained after applying the proposed methodology is widely described in the following in terms of statistics and output format.

Moreover, it is also analyzed both quantitatively and qualitatively to assess the naturalness of its utterances by investigating, first, the change of their *readability* from α to δ, and second, their well-formedness concerning syntactic (*grammaticality*) and semantic (*acceptability*) aspects.

Finally, its goodness is also assessed concerning the possibility of being used to train a deep learning model for CR in Italian.

5.1 Dataset description

Table 15 reports an overview of the obtained dataset δ, showing the total number of utterances (*utts*) and the impact of the linguistic refinements that affect a high percentage of utterances (about 64% as indicated by *refined utts*). Table 15 shows that most of the changes are related to pronouns. In particular, the row "*subject pronouns deleted being mentions*" indicates that the deletion of subject pronouns (9848 in total) overcomes rewriting rules, including both pronouns and adjectives (9045 in total). Adjectives are involved to a lesser extent in both rules, as seen from the last three rows of Table 15.

Concerning the linguistic rules applied to generate δ, the ones that have found most application instances are deletions, with the consequent shifts in coreference on the verb. This result is reasonably expected since there is a transition from a language with a mandatory expressed subject to a pro-drop language in which the subject pronoun is systematically missing.

As already mentioned, this has been one of the most challenging tasks both from a theoretical and practical point of view. The transition from a language with an explicit subject (English) to a pro-drop language (Italian) is not limited to only a deletion process. In fact, it is widespread for the subject pronoun to be labeled as mention in the original dataset, so it has almost always been necessary to shift the mention without compromising the dependencies and syntactic structure of the sentence.

The dataset δ has been structured for being released in both CoNLL and JSON formats.[2] Both formats preserve morpho-grammatical information on the parts of speech of each element of the utterance. CoNLL annotation is helpful to enable the easy interface with tools and models typically used in CR (see Fig. 9).

[2] Dataset will be made available upon request at https://nlpit.na.icar.cnr.it/nlp4it.

Table 14 The criteria followed for rewriting demonstrative pronouns

IF $t_i \in t(u)$ is	and $t_j \in t(u)$ is	and	THEN
Demonstrative Pronoun	Noun	$\nexists\, d(t_j,t_i) \in DT(u)$: $label(d) = determiner \wedge i < j \leq i + 4$	t_i is *"ciò"*

JSON version of the dataset is enriched with additional information. As shown in Fig. 10 it keeps track of changes involving utterances and mentions in the generation of the dataset δ, highlighting the data impacted from the subjects deletion, pronouns and adjective rewrite, and mentions shift.

For instance, the original utterance shown in Fig. 10 is "Esso era facile da gestire una volta che tutti capivano" (*It was easy to manage once everybody understood*), whereas it is modified by a deletion rule as shown in "*modified text.*" The rewritten utterance has a readability score equal to 79.26, it drops the pronoun subject "Esso" (*It*) with a shift of the mention from "Esso," as can be seen in "*corefs old,*" to the verb "era" (*was*). Since there is a deletion, indices indicating the position of the mention ("*start*" and "*end*") remain unchanged because the verb takes the position of the deleted subject pronoun. Notice that, this shifting process that moves the coref to the verb is consistent with the linguistic theory. The centering role of the verbal phrase within the sentence reflects theoretical aspects inherent in the hierarchical dependencies of the sentence constituents.

5.2 Readability assessment

The first evaluation of the resulting dataset δ is performed quantitatively concerning the criterion of *readability*.

In natural language, readability is defined as the ease with which a reader can understand a written text. It depends on lexical (i.e., the complexity of the vocabulary used) and syntactic factors (i.e., the presence of nested subordinate clauses). Several readability scores exist in the literature, which provides a way to assess a written text's quality automatically.

To this aim, for this work, both the readability scores based on Flesch-Vacca index [16] and the Flesch reading ease test, adapted to the Italian language, have been calculated for the utterances within both the datasets α and δ. Table 16 shows the readability scores for the dataset α (in English) and the dataset δ (in Italian). The percentage of utterances falling into each readability range are presented in each row.

The table shows that the dataset δ, expressed in the target language, is comparable to the readability of the dataset α, expressed in the source language. This result suggests that the proposed methodology has not significantly altered the overall readability of the utterances. Instead, there is singular progress in the class grouping utterances with scores above 80 (an improvement of 4.6 percentage points). As can be noted, there is a significant drop in inconsistency in judging sentences with readability between 40 and 60. This result is an expected outcome since the greater the readability, the greater is the agreement between the annotators [59].

However, even if this readability assessment gives a rough idea of the validity of the proposed methodology, it is not without limitations since the used readability scores are still debated in the literature [60]. For instance, polysyllabic words significantly affect the score, and the metrics are unbalanced on the lexicon compared to the syntax. Furthermore, a readable utterance is characterized by a linear syntax and simple vocabulary, but it can contain infelicities that make it ill-formed.

Table 15 The impact of the linguistic refinements over the dataset δ

	Train	Test	Dev
utts	44,073	5415	5363
Refined utts	28,216	3512	3471
Subject pronouns	34,974	3904	3893
Subject pronouns being mentions	14,511	1871	1517
Subject pronouns deleted being mentions	8764	1111	973
Pronouns and adjectives	37,611	6816	6728
Pronouns and adjectives being mentions	14,623	1887	1528
Pronouns and adjectives rewritten	7346	866	853

5.3 Grammaticality and acceptability assessment

The second evaluation of the resulting dataset δ is performed qualitatively to overcome the limitations of these readability scores by considering the criteria of *grammaticality* and *acceptability*.

These criteria have a long history in theoretical linguistics [61]. In detail, grammaticality refers to correct utterances from a syntactic and structural point of view according to the annotator's judgments; on the contrary, acceptability assesses whether an utterance is semantically valid according to the annotator's conclusions. In other words, grammaticality is not necessarily associated with semantic correctness or acceptability. Still, it refers to a well-formed utterance, i.e., which conforms to Italian

Fig. 9 Example of CoNLL format

```
bc/cctv/00/cctv_0005 5 0 era AUX (TOP - - - Wang_shilin * (121)
bc/cctv/00/cctv_0005 5 1 facile X * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 2 da ADP * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 3 gestire AUX * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 4 una DET * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 5 volta NOUN * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 6 che SCONJ * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 7 tutti PRON * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 8 capivano AUX * - - - Wang_shilin *
bc/cctv/00/cctv_0005 5 9 . PUNCT ) - - - Wang_shilin * -
```

Fig. 10 Example of JSON format


```
{
 "id": 26,
 "text": "Esso era facile da gestire una volta che tutti capivano .",
 "modified_flag": true,
 "modified_text": "era facile da gestire una volta che tutti capivano
 ,→ .",
 "readability": "79.26",
 "doc": "bc/cctv/00/cctv_0005",
 "part": "005",
 "speakers": "Wang_shilin",
 "tokens": ["era", "facile", "da", "gestire", "una", "volta", "che", ,→
 "tutti", "capivano", "."],
 "tags": ["AUX", "X", "ADP", "AUX", "DET", "NOUN", "SCONJ", "PRON",
 ,→ "AUX", "PUNCT" ],
 "corefs": [{ "label": "121", "text": "era", "start": 0, "end": 0 } ],
 "corefs_old": [{ "label": "121", "text": "Esso", "start": 0, "end": 0
 ,→ } ]
}
```


grammar rules. By contrast, acceptability may consider some aspects that can only be inferred by a native speaker, such as cohesion or the naturalness of the utterance.

Therefore, an utterance may be perfectly valid from a structural point of view but not be semantically comprehensible. As an example, the utterance *"Major League Baseball ha preso 76 dei suoi pipistrelli e li ha radiografati per il sughero ."* (*"Major League Baseball has taken 76 of its bats and X-rayed them for corkage ."*) is grammatical because all the constituents are in the right place, and it does not violate structural constraints. Still, there is no native speaker who would perceive it as meaningful. In this case, the error lies in translating specialized terms related to the sports domain. In particular, the term *"bat"* is ambiguous because it can refer either to the object (wooden club used in the sport of baseball to hit the ball) or to the animal (as in the incorrect translation *"pipistrello"*).

The second assessment concerning these two criteria has been carried out by considering a sample of 1000 instances extracted from the dataset δ, with 200 utterances for each readability class reported in Table 16. The extraction has not been performed entirely randomly, but it has been guided by a nonprobability sampling, which is more suitable for qualitative data. The assessment has involved three human native speakers who were asked to manually and independently label that sample by specifying, for each utterance, both its grammaticality and acceptability.

The overall agreement between these three raters concerning their annotations of grammaticality and acceptability has been measured using the Observed Agreement index [62]. This index gives a good approximation of annotators' agreement in contexts with many annotators, also offering robustness against imperfect (textual) data [63]. The index calculates the number of generated utterances with the majority agreement and reports that number as a percentage of the total number of utterances extracted by all the annotators. Grammaticality and acceptability have been calculated using forced-choice binary task [64], following most of the linguistic methodology in this area [65].

Table 17 shows the percentage of agreement between annotators for each readability class.

The total agreement value has been measured as equal to 0.78 and 0.73 in the case of annotations represented by grammatical or acceptable utterances. According to the grid for the interpretation of the coefficients proposed by [66], the values obtained indicate "substantial agreement" concerning both grammaticality and acceptability.

Human's judgements seem to be consistent with the readability scores; a higher value corresponds to a better agreement, thus a lower presence of ill-formed utterances. The agreement among the raters regarding grammaticality increases progressively (from 0.77 to 0.80) concerning the readability classes. This phenomenon can be explained by

Table 16 Comparison of readability scores before and after linguistic refinement

Score	Sentence percentage		Description
	α	δ	
> 80	38.9	43.5	Very easy to read
80–60	33.7	36.05	Fairly easy to read
60–40	17.4	15.4	Fairly difficult to read
40–20	7.2	3.8	Difficult to read
< 20	2.06	1.1	Extremely difficult to read

the fact that readability is essentially based on the utterance structure, i.e., syntax, which is the object of the grammaticality judgement.

The situation is not different as regards acceptability. First, an unsurprising slight worsening of the scores concerning lower classes has been highlighted. Lower agreement between annotators is quite common, especially in the semantic tasks [67]. However, moving on to the classes containing the most readable utterances, the values are comparable to the ones of grammaticality. Utterances considered the most readable are also those that create the slightest disagreement among the annotators, with the highest percentage of acceptable utterances.

In summarizing, the performed grammaticality and acceptability assessment has shown that the proposed methodology can generate utterances that respect a syntactic well-formedness and are perceived as natural by native speakers with a good level of agreement. The use of linguistic refinement rules helps reduce phenomena that could affect grammatical constraints (as in the case of rewrite rules) or a perceived naturalness of the sentence (as in the case of *null-subject*).

5.4 Linguistic and qualitative assessment

As further evaluation aspects, first, factors other than readability and annotators' judgements have been considered.

Utterances contained in the sample have been analyzed using different levels of linguistic analysis that include lexical, morphological and syntactic features. Considered factors range from lexical richness to the complexity of the periods, subordinates' presence, and the vocabulary used. They are summarized in Table 18. Values in Table 18 show that syntactic complexity goes through a progressive simplification from the class comprising the least readable sentences (< 20) to the most readable ones (> 80). Sentences are shorter (they move from an average length of 12.9 tokens to 7.9), and subordinating conjunctions are halved to the benefit of increased coordinating ones.

Second, this trend of syntactic and lexical simplification has been visually inspected by qualitatively examining some examples extracted from the dataset.

Table 19 collects a set of utterances for each readability class. The table is structured to visualize all possible combinations of raters' judgements. The first column is dedicated to different readability classes; the second one shows the *id* of each utterance. After that, two columns indicate if the utterance has been evaluated as grammatical (G) or acceptable (A) by human raters. Examples in Table 19 show that utterances in the less readable classes tend toward hypotaxis, with the presence of various types of subordinate clauses, whereas high-readable classes prefer elementary one-verb sentences. This outcome occurs in both well-formed and ill-formed utterances.

For instance, the utterance having *Id = 1d*, *"Il governo degli Stati Uniti pensa che i radicali, commentatori anti-americani e religiosi sono diventati ospiti frequenti all'emittente televisiva al Jazeera ."* (*"The US government believes that radical, anti-American and religious commentators have become frequent guests at the Al Jazeera television station ."*) has a readability score lower than 20, so it is challenging to read, but it is perceived as grammatical and acceptable by raters, even if it has a subordinate clause introduced by *"che"* (*"that"*), a long-distance dependency between the singular masculine noun *"commentatori"* (*"West"*) and the noun with the role of subject predicate *"ospiti"* (*"guests"*).

A similar syntactic structure is provided by the utterance in the class 20–40 having *Id = 2a*, *"Quindi Michelle le autorit'a davvero credere che questo testimone per quanto riguarda la discarica credibile , non essi?"* (*"So Michelle, do the authorities really to think this witness regarding the landfill [is] credible, not they?"*), which is full of errors that make it ungrammatical and difficult for a native speaker to understand. In detail, the verb appears in its infinitive form *"credere"* (*"to think"*) and it is not inflected in agreement with the subject noun *"autorit'a"* (*"authorities"*). Moreover, there is no verb connected to the subject complement *"credibile"* (*"credible"*) and there is a noun

Table 17 Annotator agreement for different readability classes

	< 20	20–40	40–60	60–80	> 80	Total
Grammaticality (%)	0.77	0.78	0.70	0.80	0.80	0.78
Acceptability (%)	0.64	0.64	0.75	0.83	0.81	0.73

Table 18 Different features affecting the readability on the sample considered

		Classes				
		< 20	20–40	40–60	60–80	> 80
Lexical features	Average length (tokens)	12.9	13.7072	12.1082	114.072	7.9
	Type-token ratio	0.8	0.545	0.531	0.513	0.65
	Lexical density	0.585	0.545	0.531	0.513	0.536
	Nouns	16.30%	15.10%	12.40%	13.40%	10.10%
	Proper nouns	5.60%	6.10%	5.00%	6.00%	5.30%
Morphologic features	Adjectives	5.90%	5.70%	5.20%	3.70%	3.90%
	Verbs	18.00%	20.20%	22.40%	20.20%	22.20%
	Conjunctions	4.30%	4.80%	5.50%	4.70%	5.80%
	Coordinating conjunctions	59.60%	60.70%	52.20%	69.20%	76.10%
	Subordinating conjunctions	40.40%	39.30%	47.80%	30.80%	23.90%
	Average number of clauses per utterance	1.816	1.985	2.01	1.723	1.507
	Independent clauses	71.20%	69.10%	68.20%	76.90%	92.20%
	Subordinate clauses	28.80%	30.90%	31.80%	23.10%	7.80%
Syntactic feature	Average word Number per clause	7.104	6.897	6.007	6.603	5.215
	Average DPT depth	4.595	4.813	4.456	4.436	3.015
	Average depth of noun phrase	1.133	1.131	1.134	1.142	1.064
	Average depth of subordinate chain	1.345	1.29	1.265	1.115	1.167
	Average length of dependency relations	1.913	1.906	1.848	1.77	1.795

phrase "non essi" (*non they*) at the end of the utterance completely disconnected from the syntactic structure.

In readability classes, subordinate clauses are reduced, and correlation prevails in the syntactic structure. However, this syntactic simplification does not necessarily correspond to greater comprehensibility. As mentioned above, readability tests evaluate the complexity of the lexicon and structure of the utterance. Shorter utterances having no subordinates are not always what human raters consider semantically meaningful or grammatically correct.

For instance, the utterances having *id = 5c*, *"Poi i seguaci torn'o a casa"* (*"Then the followers has gone home"*), and *id=5d*, *"La grazia di Dio sia con te"* (*"God's grace be with you"*) have a similar one-verb structure without any type of syntactic or lexical complexity. However, the utterance *5c* contains a grammatical infelicity, with a wrong agreement between the 3rd person singular verb *"torn'o"* (*"has gone"*) and the plural subject noun *"seguaci"* (*"followers"*).

In summary, readability scores obtained automatically have proven to be consistent with the raters' judgments, allowing sentences to be grouped into classes that are in line with grammaticality and acceptability. However, it should be noted that there are numerous other linguistic variables affecting readability that are independent of the metrics used, but this is outside the scope of this work.

5.5 Effectiveness assessment as training dataset

The last evaluation has been performed to assess the goodness of the generated dataset concerning the possibility of being used to train a deep learning model for CR in Italian. To this aim, a baseline model on the dataset is generated by adopting a state-of-the-art deep learning architecture proposed for the same task in English. In detail, the coreference model proposed by [44] has been used,[3] by exploiting BERT in its base (cased) version.[4]

This choice is justified by the fact that this model has proven to be effective in the CR task in English, as shown in [68–70].

To the best of our knowledge, no other available implementation exists for the particular CR task in Italian.

In detail, the architecture of BERT is characterized by 12 encoder layers, known as Transformers Blocks, and 12 attention heads (or Self-Attention as introduced in [71]), hence feedforward networks with a hidden size of 768. Each training session has been fixed of 24 epochs, with a variable learning rate from 0.1 to 0.00001. More architectural details and training hyperparameters are reported in Table 20. All experiments have been performed on a deep learning workstation, with 40 Intel(R) Xeon(R) CPUs

[3] https://github.com/lxucs/coref-hoi.

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

Table 19 Visual examples of sentences for each readability class and raters' judgements

Class	Id	G	A	Utterance
< 20	*1a*	−	+	Quattro esplosioni strappare attraverso la metropolitana vedere
				(Four explosions rip through the metro see)
	1b	−	+	Deputati dell'opposizione stanno esprimendo suo malcontento
				(Opposition deputies are expressing his dissatisfaction.)
	1c	+	−	Occasionalmente, il danno cromosoma lordo era visibile
				(Occasionally, gross chromosome damage was visible.)
	1d	+	+	Il governo degli Stati Uniti pensa che i radicali, commentatori antiamericani e religiosi sono diventati ospiti frequenti all'emittente televisiva al Jazeera
				(The US government believes that radical, anti-American and religious commentators have become frequent guests at the al Jazeera television station.)
20–40	*2a*	−	−	Quindi Michelle le autorita' davvero credere che questo testimone per quanto riguarda la discarica credibile, non essi ?
				(So Michelle, do the authorities really think this witness regarding the landfill is credible?)
	2b	−	+	Essi ha risposto, Si, Signore, crediamo
				(They replied, Yes, Lord, we believe.)
	2c	+	−	riadattamento per quanto riguarda gli americanismi
				(readjustment with regard to Americanisms)
	2d	+	+	chiaramente crediamo che Davis stava resistendo
				(We clearly believe that Davis was resisting.)
40–60	*3a*	−	−	sua politica sono incorporati nella scrittura Di suo e suo scrivere è prima di tutto una celebrazione della libertà
				(his politics are embedded in his writing and his writing is first and foremost a celebration of freedom)
	3b	−	+	Poi trascorrere piu' tempo con Loro, e incoraggiare Loro per ottenere piu' esercizio fisico e prendersi cura di Stessi
				(Then spend more time with them and encourage them to get more exercise and take care of themselves.)
	3c	+	−	e che include Cinquanta centesimi troppo
				(and that includes Fifty Cents Too)
	3d	+	+	In primo luogo, Stoccolma ha speso 180 milioni di dollari per i miglioramenti dei trasporti prima dell' l'esperimento
				(First, Stockholm spent 180 million dollars on transport improvements before the experiment)
60–80	*4a*	−	−	Il video suona anche le voci di coloro che si accanto al cadavere parlando tra loro
				(The video also plays the voices of those next to the corpse talking to each other)
	4b	−	+	Avranno bisogno di un' enorme somma di denaro per portare i bambini Loro in città
				(They will need a huge amount of money to bring their children to the city.)
	4c	+	−	Ho chiesto a i tuoi seguaci di forzare lo spirito malvagio fuori
				(I have asked your followers to force the evil spirit out.)
	4d	+	+	Questo è l' insegnamento che avete sempre sentito: dobbiamo amarci l' un l' altro
				(This is the lesson you have always heard: we must love one another.)
> 80	*5a*	−	−	Lui ha messo le mani di suo su Suo, e subito Lei è riuscita a stare dritta. (He put his hands on hers, and immediately she managed to stand up straight.)
	5b	−	+	la cosa con il Golan per dare esso indietro di non dare esso indietro non lo so
				(the thing with the Golan to give it back not to give it back I do not know.)
	5c	+	−	Poi i seguaci torn'o a casa. (Then the followers went home.)
	5d	+	+	La grazia di Dio sia con te. (God's grace be with you)

E5-2630 v4 @ 2.20 GHz, 256 GB of RAM and 4 GPUs GeForce GTX 1080 Ti. The operating system is Ubuntu Linux 16.04.7 LTS. Using the train division of the created dataset, the results have been derived by averaging the performance of the coreference model over five repetitions and finally reporting the arithmetic mean of the results, rounded to the second decimal place. Table 21 reports the results obtained with three different metrics: *MUC* [72], B^3 [73] and $CEAF_{\phi 4}$ [74].

MUC provides a good measure of the interpretability achieved by the model, which indicates the goodness in the prediction of mentions and coreference links among them.

However, *MUC* lacks discriminability, i.e., the capability to distinguish between good and not good decisions. On the contrary, B^3 and $CEAF_{\phi 4}$ lack interpretability, but they measure discriminability. Since none of the metrics is reliable if taken individually, it is common practice to use the average of the three as the overall metric.

As shown in Table 21, *MUC* has achieved better performances on precision and recall, respectively. $CEAF_{\phi 4}$, instead, has the lowest scores, especially concerning recall (about 59.25). B^3 provides scores quite similar to those obtained with $CEAF_{\phi 4}$. On average, the model has achieved an F1 of about 69,60, which is comparable with the averaged F1 obtained by the same model but on the English version of the Ontonotes dataset (about 73.9).

As an example, sentences extracted from the dataset and shown in Table 22 present cases of correct mention predictions and wrong ones. Predictions are indicated in bold, while mentions to which the predictions refer are shown in small caps.

Concerning the analysis on the typology of errors, in the first sentence "[Essi] hanno scritto oggi" (*They wrote today*) the correctly predicted mention occurs as a verb in the English text, and it has been shifted on the verb in the Italian one due to the drop of the subject pronoun "Essi" (*They*). The second example presents a linear subject–verb–object sentence with an explicit subject. In this case, the proper noun acting as a subject is into a prepositional phrase "L'ex avvocato di Clinton" (*Clinton's former lawyer*), and it is correctly predicted. Moving to the analysis of incorrectly recognized predictions, it is possible to note that a more complex syntax affects the predictions. For instance, in the first example (first sentence of the wrong predicted row) the utterance contains a dative construction with a clitic pronoun "Ci" (literally *us*) preceding the mention "riferivamo" (*were referring*) and an enclitic form merged with the verb in the form of suffix -lo for the coreference "farlo" (*to do that*). Finally, in the last example, BERT fails the correct assignment when the mention occurs as indirect object introduced by a preposition "a questo" (*about this*).

In spite of special cases such as those described above (clitics, convoluted syntax), these results have shown the effectiveness of the proposed methodology, providing a new dataset for CR in Italian and setting a baseline for future developments of this line of research.

6 Conclusions and future work

This work presents a methodology for creating a dataset for CR in Italian starting from a resource initially designed for English. This approach can guarantee a quality comparable to manual annotation while reducing the time and effort it requires. Starting from the OntoNotes, this methodology has been articulated in two macro-steps.

Table 20 Hyper-parameters

Hyperparameter	Value
Epochs	24
Dropout	0.3
Learning rate	From 0.1 up to 0.00001
Loss	Marginalized
Feature embedding size	20
Max span width	30
Max training sentences	6
Max segment length	256
Dimensions hidden state	256
Number of attention heads	12
Number of hidden layers	12
Hidden size	768
Number of hidden layers	12
Parameters	110 M
Vocabulary size	32,102

The first macro-step is focused on the generation of a corpus in the target language. This step first extracts from the OntoNotes the information of interest, such as documents, partitions, utterances, and mentions, but discharging irrelevant information and mentions whose tokens are contained in other mentions. Then, utterances and mentions are translated through an intelligent token replacement/resolution procedure guided by the estimation of the typology, gender and number of the real-world entities referred by each mention. The second macro-step is focused on linguistic refinement. This step first tries to correct all the infelicities introduced in the translation on aspects of the Italian language not present in English (i.e., gender and number agreement). Then, it attempts to make translated utterances more natural as perceived by a native speaker (*null-subject*).

The well-formedness and naturalness of the generated dataset has been confirmed by means of a quantitative and qualitative assessment, which has evaluated readability on all the utterances of the final dataset and grammaticality and acceptability on a sample of 1000 utterances extracted from different five readability classes by three human native speakers. A correlation between the readability score and raters' judgements has been also highlighted, with utterances featuring poor readability having the highest disagreement among human raters for both grammaticality and acceptability. The goodness of the dataset has also been assessed by training a CR model based on BERT,

Table 21 Results achieved with a BERT-based CR model

MUC			*B3*			$CEAF_{\phi 4}$			avg *F*1
R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	
73.44	79.56	76.38	64.19	70.83	67.34	59.25	72.24	65.10	69.60

Table 22 Examples of correct and wrong predictions (bold) with respect to mentions (small caps)

Correctly predicted	HANNO SCRITTO oggi… Lei non ha condiviso le note con **loro**
	(THEY wrote today… She did not share the notes with **them**)
	L'ex avvocato di CLINTON… Sono mosse accuse contro di **lui**
	(CLINTON's former lawyer… Allegations are made against **him**)
Wrong predicted	CI RIFERIVAMO a esso… è sempre difficile far**lo**
	(WE were referring to it… it is always difficult to do **that**)
	Ho pensato a lungo a QUESTO… Molti criticano **ciò**
	(I thought about this for a long time… Many people criticise **this**)

In brackets the English text

achieving promising results and thus, fixing a reference point in terms of performance for future comparisons.

It is worth noting that, for this work, English has been considered as the source language and Italian as the target one, due to the high and limited number of existing resources existing for them, respectively. However, the methodology is not strictly dependent on these two languages and can be easily applied to other languages, by only adapting a small set of linguistic rules.

From a methodological perspective, even if the quality of the final dataset is appreciable, it leaves room for some future improvements. First, a more extensive list of refinement rules regarding other linguistic phenomena of the Italian language will be considered to enhance the naturalness of the translated utterances. Second, utterances with more complex syntactic structures will be handled to improve readability, grammaticality and acceptability. From an applicative perspective, the dataset will be used to train novel and better performing models for the task of CR in Italian.

Data availability The dataset described in this study will be available at the address https://nlpit.na.icar.cnr.it/nlp4it/#/datasets/.

Declaration

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

References

1. Sukthanker R, Poria S, Cambria E, Thirunavukarasu R (2020) Anaphora and coreference resolution: a review. Inform Fusion 59:139–162
2. Antunes J, Lins RD, Lima R, Oliveira H, Riss M, Simske SJ (2018) Automatic cohesive summarization with pronominal anaphora resolution. Comput Speech Lang 52:141–164
3. Sikdar UK, Ekbal A, Saha S (2016) A generalized framework for anaphora resolution in Indian languages. Knowl Based Syst 109:147–159
4. Blackwell SE (2001) Testing the Neo-Gricean pragmatic theory of anaphora: the influence of consistency constraints on interpretations of coreference in Spanish. J Pragmat 33(6):901–941
5. Lee C, Jung S, Park C-E (2017) Anaphora resolution with pointer networks. Pattern Recogn Lett 95:1–7
6. Stylianou N, Vlahavas I (2021) A neural entity coreference resolution review. Expert Syst Appl 168:114466
7. Clark K, Manning CD (2016) Deep reinforcement learning for mentionranking coreference models. arXiv preprint arXiv:1609.08667
8. Zheng J, Chapman WW, Crowley RS, Savova GK (2011) Coreference resolution: a review of general methodologies and applications in the clinical domain. J Biomed Inform 44(6):1113–1122
9. Hirschman L, Chinchor N (1997) Muc-7 proceedings. Science Applications International Corporation. See www.muc.saic.com
10. Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y (2012) Conll-2012 shared task: modeling multilingual unrestricted coreference in ontonotes. In: Joint conference on EMNLP and CoNLL-shared task, pp 1–40

11. Recasens M, Hovy E (2011) Blanc: Implementing the rand index for coreference evaluation. Nat Lang Eng 17(4):485–510
12. Poesio M, Delmonte R, Bristot A, Chiran L, Tonelli S (2004) The Venex corpus of anaphora and deixis In spoken and written Italian. University of Essex
13. Magnini B, Pianta E, Girardi C, Negri M, Romano L, Speranza M, Bartalesi V, Sprugnoli R (2006) I-cab: the Italian content annotation bank. In: 5th International conference on language resources and evaluation (LREC 2006), pp 963–968
14. Rodrıguez KJ, Delogu F, Versley Y, Stemle EW, Poesio M (2010) Anaphoric annotation of Wikipedia and blogs in the live memories corpus. In: Proceedings of LREC, pp 157–163
15. Hovy E, Marcus M, Palmer M, Ramshaw L, Weischedel R (2006) Ontonotes: the 90% solution. In: Proceedings of the human language technology conference of the NAACL, companion volume: short papers, pp 57–60
16. Franchina V, Vacca R (1986) Adaptation of flesh readability index on a bilingual text written by the same author both in Italian and English languages. Linguaggi 3:47–49
17. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
18. Pradhan SS, Ramshaw L, Weischedel R, MacBride J, Micciulla L (2007) Unrestricted coreference: identifying entities and events in ontonotes. In: International conference on semantic computing (ICSC 2007). IEEE, pp 446–453
19. Grishman R, Sundheim BM (1996) Message understanding conference-6: a brief history. In: COLING 1996 volume 1: The 16th international conference on computational linguistics
20. Chinchor NA (1998) Overview of muc-7/met-2. Technical report, Science Applications International Corp San Diego
21. Poesio M (2004) Discourse annotation and semantic annotation in the gnome corpus. In: Proceedings of the workshop on discourse annotation, pp 72–79
22. Poesio M, Artstein R et al (2008) Anaphoric annotation in the Arrau corpus. In: LREC
23. Chen YH, Choi JD (2016) Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In: Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue, pp 90–100
24. Cybulska A, Vossen P (2014) Guidelines for ecb+ annotation of events and their coreference. In: Technical report NWR-2014-1, VU University Amsterdam
25. Zeldes A, Zhang S (2016) When annotation schemes change rules help: a configurable approach to coreference resolution beyond ontonotes. In: Proceedings of the workshop on coreference resolution beyond OntoNotes (CORBON 2016), pp 92–101
26. Ghaddar A, Langlais P (2016) Wikicoref: an English coreference-annotated corpus of wikipedia articles. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 136–142
27. Marcus MP, Marcinkiewicz MA (2004) Building a large annotated corpus of English: the penn treebank. Comput Linguist 19(2)
28. Hasler L, Orasan C, Naumann K (2006) Nps for events: experiments in coreference annotation. In: Proceedings of the fifth international conference on language resources and evaluation (LREC'06)
29. Kim J-D, Ohta T, Tateisi Y, Tsujii J (2003) Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19(suppl 1):180–182
30. Tateisi Y, Yakushiji A, Ohta T, Tsujii J (2005) Syntax annotation for the Genia corpus. In: Companion volume to the proceedings of conference including posters/demos and tutorial abstracts
31. Kim J-D, Ohta T, Tsujii J (2008) Corpus annotation for mining biomedical events from literature. BMC Bioinform 9(1):10
32. Su J, Yang X, Hong H, Tateisi Y, Tsujii J (2008) Coreference resolution in biomedical texts: a machine learning approach. In: Dagstuhl seminar proceedings. Schloss Dagstuhl-Leibniz-Zentrum für Informatik
33. Nguyen TORBN, Kim JTJD, Pyysalo S (2011) Overview of bionlp shared task 2011. In: Proceedings of BioNLP shared task 2011 workshop, pp 1–6
34. Cohen KB, Johnson HL, Verspoor K, Roeder C, Hunter LE (2010) The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinform 11(1):492
35. Batista-Navarro RT, Ananiadou S (2011) Building a coreference-annotated corpus from the domain of biochemistry. In: Proceedings of BioNLP 2011 workshop, pp 83–91
36. Segura-Bedmar I, Crespo M, de Pablo C, Martınez P (2009) Drugnerar: linguistic rule-based anaphora resolver for drug-drug interaction extraction in pharmacological documents. In: Proceedings of the third international workshop on data and text mining in bioinformatics, pp 19–26
37. Doddington GR, Mitchell A, Przybocki MA, Ramshaw LA, Strassel SM, Weischedel RM (2004) The automatic content extraction (ace) program-tasks, data, and evaluation. In: Lrec, vol 2. Lisbon, pp 837–840
38. Weischedel R, Palmer M, Marcus M, Hovy E, Pradhan S, Ramshaw L, Xue N, Taylor A, Kaufman J, Franchini M et al (2013) Ontonotes release 5.0 ldc2013t19. Linguistic Data Consortium, Philadelphia, p 23
39. Recasens M, Marquez L, Sapena E, Martı MA, Taule M, Hoste V, Poesio M, Versley Y (2010) Semeval-2010 task 1: coreference resolution in multiple languages. In: Proceedings of the 5th international workshop on semantic evaluation, pp 1–8
40. Guillou L, Hardmeier C, Smith A, Tiedemann J, Webber B (2014) Parcor 1.0: a parallel pronoun-coreference corpus to support statistical mt. In: 9th International conference on language resources and evaluation (LREC), May 26–31, 2014, Reykjavik, ICELAND. European Language Resources Association, pp 3191–3198
41. Montemagni S, Barsotti F, Battista M, Calzolari N, Corazzari O, Zampolli A, Fanciulli F, Massetani M, Raffaelli R, Basili R et al (2003) The Italian syntactic-semantic treebank: architecture, annotation, tools and evaluation
42. Bristot A, Chiran L, Delmonte R (2000) Verso un'annotazione xml di dialoghi spontanei per l'analisi sintattico-semantica. XI Giornate di Studio GFS, Multimodalita'e Multimedialit nella comunicazione, pp 42–50
43. Pradhan S, Ramshaw L, Marcus M, Palmer M, Weischedel R, Xue N (2011) Conll-2011 shared task: modeling unrestricted coreference in ontonotes. In: Proceedings of the fifteenth conference on computational natural language learning: shared task, pp 1–27
44. Lee K, He L, Lewis M, Zettlemoyer L (2017) End-to-end neural coreference resolution. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 188–197
45. Lakretz Y, Hupkes D, Vergallito A, Marelli M, Baroni M, Dehaene S (2020) Exploring processing of nested dependencies in neural-network language models and humans. arXiv preprint arXiv:2006.11098
46. Kabadjov MA (2007) A comprehensive evaluation of anaphora resolution and discourse-new classification. PhD thesis, Citeseer
47. Liu H (2010) Dependency direction as a means of word-order typology: a method based on dependency treebanks. Lingua 120(6):1567–1578. https://doi.org/10.1016/j.lingua.2009.10.001
48. Tsarfaty R, Seddah D, Goldberg Y, Kuebler S, Versley Y, Candito M, Foster J, Rehbein I, Tounsi L (2010) Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In: Proceedings of the NAACL HLT 2010 first workshop on

statistical parsing of morphologically-rich languages. Association for Computational Linguistics, Los Angeles, pp 1–12. https://www.aclweb.org/anthology/W10-1401
49. Liu H, Xu C (2012) Quantitative typological analysis of Romance languages. Poznan Stud Contemp Linguist 48(4):597–625. https://doi.org/10.1515/psicl-2012-0027
50. Wang L, Tu Z, Zhang X, Liu S, Li H, Way A, Liu Q (2017) A novel and robust approach for pro-drop language translation. Mach Transl 31(1–2):65–87
51. Wang L, Tu Z, Shi S, Zhang T, Graham Y, Liu Q (2018) Translating pro-drop languages with reconstruction models. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI18). AAAI Press, New Orleans, pp 4937–4945. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16187
52. Evans R (2001) Applying machine learning toward an automatic classification of it. Literary Linguist Comput 16(1):45–58
53. Yin Q, Zhang Y, Zhang W, Liu T, Wang WY (2018) Zero pronoun resolution with attention-based neural network. In: Proceedings of the 27th international conference on computational linguistics, pp 13–23
54. Gopal M, Jha GN (2017) Zero pronouns and their resolution in Sanskrit texts. In: The international symposium on intelligent systems technologies and applications. Springer, pp 255–267
55. Aloraini A, Poesio M et al (2020) Cross-lingual zero pronoun resolution
56. Guarasci R, Silvestri S, De Pietro G, Fujita H, Esposito M (2022) Bert syntactic transfer: a computational experiment on Italian, French and English languages. Comput Speech Lang 71:101261
57. McKelvie D, Isard A, Mengel A, Baun Møller M, Grosse M, Klein M (2001) The mate workbench—an annotation tool for xml coded speech corpora. Speech Commun 33(1):97–112. https://doi.org/10.1016/S0167-6393(00)00071-6
58. Lakretz Y, Dehaene S, King J-R (2020) What limits our capacity to process nested long-range dependencies in sentence comprehension? Entropy 22(4):446
59. Dell'Orletta F, Wieling M, Venturi G, Cimino A, Montemagni S (2014) Assessing the readability of sentences: which corpora and features? In: Proceedings of the ninth workshop on innovative use of NLP for building educational applications, pp 163–173
60. Crossley SA, Skalicky S, Dascalu M, McNamara DS, Kyle K (2017) Predicting text comprehension, processing, and familiarity in adult readers: new approaches to readability formulas. Discourse Process 54(5–6):340–359
61. Sprouse J (2018) Acceptability judgments and grammaticality, prospects and challenges. Syntactic structures after 60 years: the impact of the Chomskyan revolution in linguistics, vol 129, pp 195–224
62. Kruskal WH, Goodman L (1954) Measures of association for cross classifications. J Am Stat Assoc 49(268):732–764
63. Bobicev V, Sokolova M (2017) Inter-annotator agreement in sentiment analysis: machine learning perspective. In: RANLP, pp 97–102
64. Sprouse J, Schutze CT, Almeida D (2013) A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. Lingua 134:219–248. https://doi.org/10.1016/j.lingua.2013.07.002
65. Langsford S, Perfors A, Hendrickson AT, Kennedy LA, Navarro DJ (2018) Quantifying sentence acceptability measures: reliability, bias, and variability. Glossa J Gen Linguist 3(1):37. https://doi.org/10.5334/gjgl.396
66. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 159–174
67. Aroyo L, Welty C (2015) Truth is a lie: crowd truth and the seven myths of human annotation. AI Mag 36(1):15–24
68. Joshi M, Levy O, Zettlemoyer L, Weld D (2019) BERT for coreference resolution: baselines and analysis. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 5803–5808. https://doi.org/10.18653/v1/D19-1588
69. Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) Spanbert: improving pre-training by representing and predicting spans. Trans Assoc Comput Linguist 8:64–77
70. Xu L, Choi JD (2020) Revealing the myth of higher-order inference in coreference resolution. arXiv preprint arXiv:2009.12013
71. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
72. Vilain M, Burger JD, Aberdeen J, Connolly D, Hirschman L (1995) A model-theoretic coreference scoring scheme. In: Sixth message understanding conference (MUC-6): proceedings of a conference held in Columbia, Maryland, November 6–8, 1995
73. Bagga A (1998) Algorithms for scoring coreference chains. In: Proceedings of linguistic coreference workshop at the first conf. on language resources and evaluation (LREC), Granada, Spain, May 1998
74. Luo X (2005) On coreference resolution performance metrics. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing, pp 25–32