Research Article

International Journal of Media and Networks

Code Book for the Annotation of Diverse Cross-Document Coreference of Entities in News Articles

Jakob Vogel*

M.A. Digital Humanities, Institute for Digital Humanities, Faculty of Philosophy, Georg August University of Gottingen

***Corresponding Author**
Jakob Vogel, M.A. Digital Humanities, Institute for Digital Humanities, Faculty of Philosophy, Georg August University of Gottingen.

Submitted: 2023, Oct 01 ; **Accepted:** 2023, Oct 25; **Published:** 2023, Nov 02

Citation: Vogel, J. (2023). Code Book for the Annotation of Diverse Cross-Document Coreference of Entities in News Articles. *Int J Med Net, 1*(1), 50-59.

Abstract

This paper presents a scheme for annotating coreference across news articles, extending beyond traditional identity relations by also considering near-identity and bridging relations. It includes a precise description of how to set up Inception, a respective annotation tool, how to annotate entities in news articles, connect them with diverse coreferential relations, and link them across documents to Wiki data's global knowledge graph. This multi-layered annotation approach is discussed in the context of the problem of media bias. Our main contribution lies in providing a methodology for creating a diverse cross-document coreference corpus which can be applied to the analysis of media bias by word-choice and labelling.

Keywords: Coreference Resolution, Diverse Coreference Annotation, Entity Annotation, Entity Linking, Media Bias Analysis, Natural Language Processing.

1. Introduction

Coreference is the phenomenon of several expressions in a text all referring to the same person, object, or other entity or event as their referent. Thus, in a narrow sense, analysing a document with regards to coreference means detecting relations of identity between phrases. The following example (1) illustrates such an identity relation, where coreferential expressions are printed in italics:

(1) *"Joe Biden* arrived in Berlin yesterday, but *the president* did not come alone."

In (1), the noun phrase "Joe Biden" introduces a new entity while *"the president"* relates back to that introducing phrase. Within this relation, the introducing phrase *"Joe Biden"* is called the antecedent while the back-relating phrase *"the president"* is called an anaphor. Both expressions are coreferential in the way that they refer to the same non-textual entity, namely to the actual 'real world' Joe Biden or at least to a corresponding mental concept. We can think of an antecedent and its anaphora as forming a cluster of mentions that as a whole represents its extra-textual referent within a textual document, as shown in Figure 1.

As a task of natural language processing (NLP), coreference resolution has become quite efficient in detecting identity relations between phrases. However, reflecting on how we use language to refer to something, we are forced to realize that coreference in a broader sense is actually far more complex. We can address an entity or event by using a variety

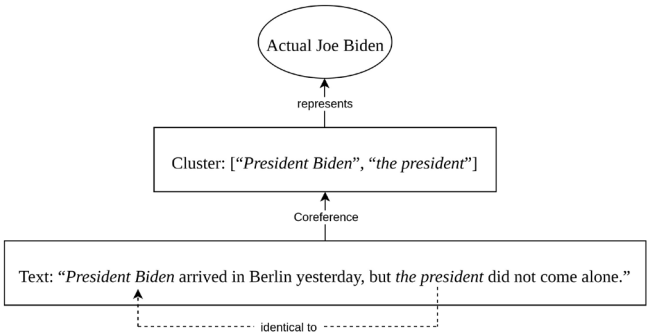

Figure 1: Illustration of How a Cluster Can be Formed From an Antecedent and its Anaphor(s). The Cluster Represents its Referent, in this Case Joe Biden, in a Text.

of expressions that are in fact not strictly identical to each other. Consider the following examples:

(2) *"President Biden* was clearly not satisfied with today's outcome. As *the White House* stated this afternoon, efforts will be made to ..."

(3) "Even if *the young Erdogan* used to be pro-Western, *Turkey's president* nowadays often acts against Western interests."

(4) "The AfD is circulating *a photo of Angela Merkel with a Hijab*, although *Merkel* never wore Muslim clothes."

In these given examples, the highlighted mentions mean 'almost' the same, but not completely. In (2), we are aware by world-

knowledge that *"the White House"* is often used as a substitute expression for the current US president, although the former is a place which in strict terms cannot be identical to the president, who is a person. In (3), on the other hand, both mentions refer to the 'real-world' person Erdogan, but at different time steps. Finally, in (4), a mention representing the person Merkel is juxtaposed with a mention representing a picture of Merkel. While these two mentions could refer to separate entities, the juxtaposition indicates a connection between both where the attributes of the first mention do influence the perception of the second mention. Hence, we would miss essential semantic connections if we chose not to mark them as coreferential. Having said that, the simple classification of two mentions into coreferential (identical) or non co referential (non-identical) does not seem to suffice the complexity of common text data. Instead, we need to allow for diverse coreference clusters that include finer-grained relations lying between identity and non-identity. We need to allow for near-identity relations to mark two mentions that are partially, but not totally, identical [1].

In news coverage, identity and near-identity references are extensively used to report on persons, organizations, and other entities of public interest. It is our goal to build up a corpus that contains annotated examples of such diverse forms of coreference. While diverse coreference occurs in all sorts of news media, we focus on digital print media, only. Furthermore, although in practice both entities and events can act as referent, we ignore references to events for now, as their annotation would go beyond the limits of our present scheme. The ordinary business of journalism is to write about current political affairs and other happenings of public interest. These happenings are normally reported by several newspapers at the same time. All of these news articles are considered documents that contain references to the same entities and together form a discourse about them. To include the whole picture of such interdiscursive references, we want our corpus to link document-level clusters with corresponding clusters of other documents of the same discourse. Hence, our corpus is to depict cross-document coreference data. On a discourse level, corresponding clusters form discourse entities that themselves can be linked to their non-textual referents by some knowledge graph identifier. For this project, we use Wiki data's Uniform Resource Identifiers (URIs) for entity linking. By doing so, world knowledge is included into the data. This allows for drawing connections even between different discourse entities that refer to a common referent, yet at a different time step or rather in the context of a different happening. Figure 2 illustrates the multiple layers of this annotation model.

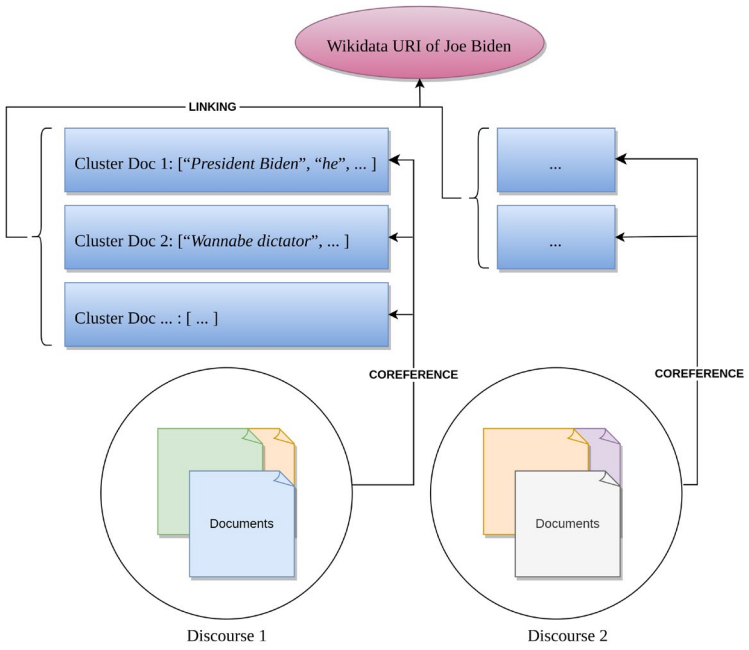

Figure 2: Illustration of our multi-layered annotation: within several discourses which all consist of multiple news articles reporting on the same happening, document entity clusters are extracted for each document. Those clusters are assigned a Wiki data URI. This ensures an unambiguous identification of each cluster, but it also links each cluster to all other clusters with the same referent within one discourse as well as across discourses. Finally, the linking also adds world knowledge to the annotated data.

In building a corpus for diverse cross-document coreference in news articles, we hope to provide a valuable resource for the evaluation of automated coreference resolution tasks. The contribution of this paper mainly lies in providing an answer to the question of how to create such a corpus. How can diverse coreference relations be annotated in a cross-document setup? We believe our scheme as we present it here tackles this problem efficiently, extensively, and unambiguously.

Additionally, we would like to use the data resulting from our own annotations for further research in the area of media bias. Even if it plays no direct part in the outlined scheme, a lot of

our choices how to annotate references were made because of this requirement to make the data usable for later media bias analysis. Eventually, we hope to contribute to the wider research question of how to identify media bias by word choice and labelling based on the usage of diverse coreference relations in news articles.

The following section 2 will further elaborate on this connection between diverse coreference and the problem of media bias analysis. Despite its only subtle impact on our practical annotation instructions, that section means to highlight the theoretical background and motivation behind our project. The sections thereafter will then deal with the actual annotation process. Section 3 will guide coders through the setup and controls of Inception, our selected annotation software. Finally, section 4 will define annotation instructions in three passes while also outlining our typology of diverse coreference.

The data we use for our own annotations consists of the text bodies of articles that report on the same happenings. All articles are in English and were published by one of the following US-American newspapers: HuffPost (categorized as ”Left” by All Sides (2023) or ”Skews Left” by Ad Fontes Media (2023), abbreviated in our data as ”LL”), The New York Times (categorized as ”Lean Left” by All Sides or ”Skews Left” by Ad Fontes Media, abbreviated as ”L”), USA Today (categorized as ”Lean Left” or ”Middle or Balanced Bias”, abbreviated as ”M”), Fox News (categorized as ”Right” or ”Skews Right”, abbreviated as ”R”), Breitbart News Network (categorized as ”Right” or ”Strong Right”, abbreviated as ”RR”).

2. Diverse Cross-Document Coreference and Media Bias Analysis

Media bias is a multifaceted phenomenon of news coverage that is one-sided, politically shaded, or in some other way non-neutral. It can occur in all sorts of news media, though we focus on digital print media, only. One specific type of media bias is bias by word-choice and labelling [2]. Word choice describes the selection from a variety of possible expressions to refer to an entity. For example, in order to refer to the USA’s current head of state, journalists could use one of the relatively neutral alternatives “Joe Biden”, “Biden”, or “the US president”, or in theory, choose a clearly biased expression like ”the dictator” [3].

Labelling, on the other hand, describes the assignment of attributes to an expression, inter alia by adding adjectives. Examples for bias by labelling include “an anxious and uncertain president” or “crooked Joe Biden” [4]. Together, word-choice and labelling form a so-called frame [2]. In news articles, frames are used in a variety of ways, either for the sake of linguistic diversity or to make certain, potentially biased statements about an entity. To test an article for such statements, all of an entity’s frames need to be extracted and evaluated together. Hence, before an article can be properly analysed with regards to if and how it uses biased frames of (certain) entities, we are first faced with the task of identifying such frames. The identification of all expressions that refer to the same entity is a matter of coreference resolution. To conclude, successful coreference resolution is a prerequisite to any further inquiry of media bias by word-choice and labelling.

As already indicated above, automatic coreference resolution does show good results in extracting identity clusters from a document [5]. However, we have seen that there exist near identity relations between expressions, potentially even across documents, that would be mostly overseen by standard coreference resolution approaches [6]. Hence, they would also be overseen by any media bias analysis that depends on coreference resolution. We hope that our building of a corpus for diverse cross-document coreference will contribute to the analysis of media bias by providing data that contains the full variety of frames used in news articles. Eventually, we would like to test how we can measure media bias by focusing on diverse coreference in news articles. To answer this last question, though, an additional layer of media bias annotation would have to be put upon our coreference data [7,8].

3. Annotation Tool

The software we will use for annotation is called Inception [9]. Inception is an open source annotation tool which can be freely downloaded from the authors’ GitHub repository. Although for this project, every annotator will be provided with a ready-to-code version of the program with all necessary annotation layers and settings already implemented and some sample annotations included. This instance of Inception can be requested from the project administrator Jakob Vogel.

3.1. Setup

To set up Inception on your local computer, make sure you already received your personal instance of the software. If not, please contact the project administrator.

Inception comes as a jar-file. In order to run it, you need to have the Java Runtime Environment (JRE) installed. Furthermore, make sure the file is set as executable. Then open the directory ”Inception” in your command prompt and run:

```
java −jarinception . jar
```

To access Inception’s graphical user interface (GUI), go to a web browser and open:
http://localhost:8080/
On your first time running Inception, you will need to import the project and set up your personal user account:
• First, log in as admin (User ID: *admin* ; Password: *admin*).
• Click on ”Import project” and select the file ”proj-div-CDCR.zip” from the ”Inception” directory. Make sure to check the boxes ”Import permissions” (already checked by default) and ”Create missing users” (unchecked by default). Then click ”Import”. • Click on ”Administration” in the GUI’s right top corner. Then click on ”Users”. • Select your personal user or create a new one here. Assign a password to your user. Additionally, assign the role ”ROLE_USER” to your user (already assigned by default). Finally, check the box ”Account enabled” and click ”Save”.

• Log Out of the Current Inception Session

From now on, to log into Inception, use your personal user account details instead of the admin account.

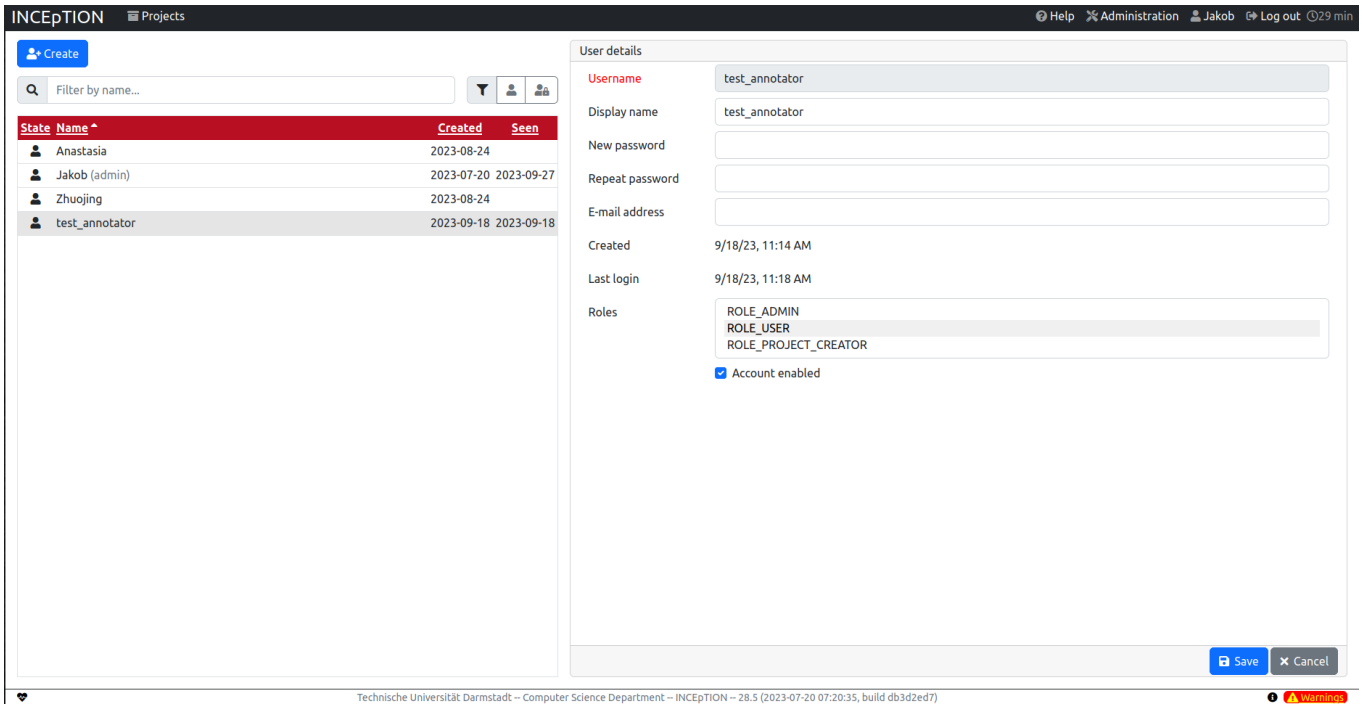

Figure 3: Screenshot of Inception Window Showing the User Management Settings. Make Sure to Create Or Activate Your Own User Account Here at First Login.

To get to the annotation GUI, log in with your personal account now and click on the highlighted project name ”Diverse cross-document coreference”. Then, in the left taskbar, click on ”Annotation”. A window opens that shows a list of all documents to be annotated. The first digit in every title is a discourse identifier that sorts all documents according to their topic, followed by an underscore and a newspaper abbreviation (see Introduction). You can annotate documents in chronological order or randomly, whichever you prefer. Click on one of the documents to start your annotation.

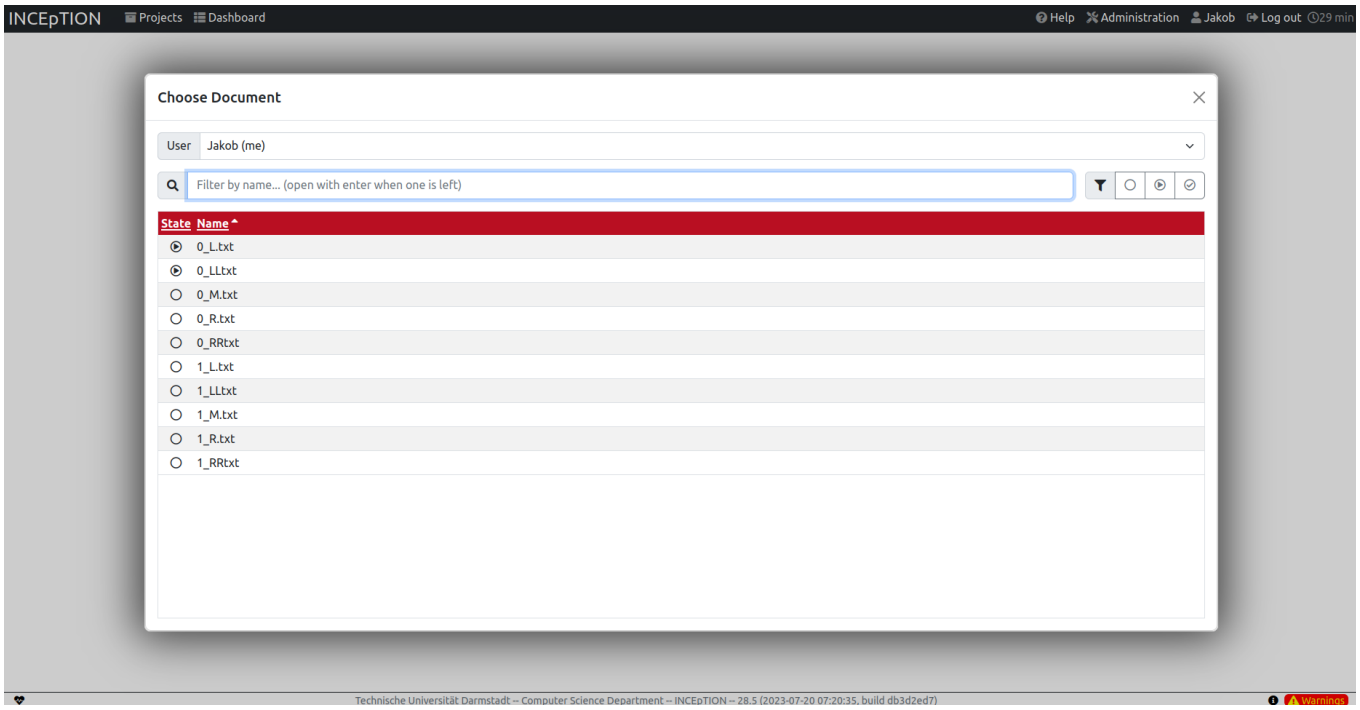

Figure 4: Screenshot of Inception Window Showing a List of All Documents to be Annotated.

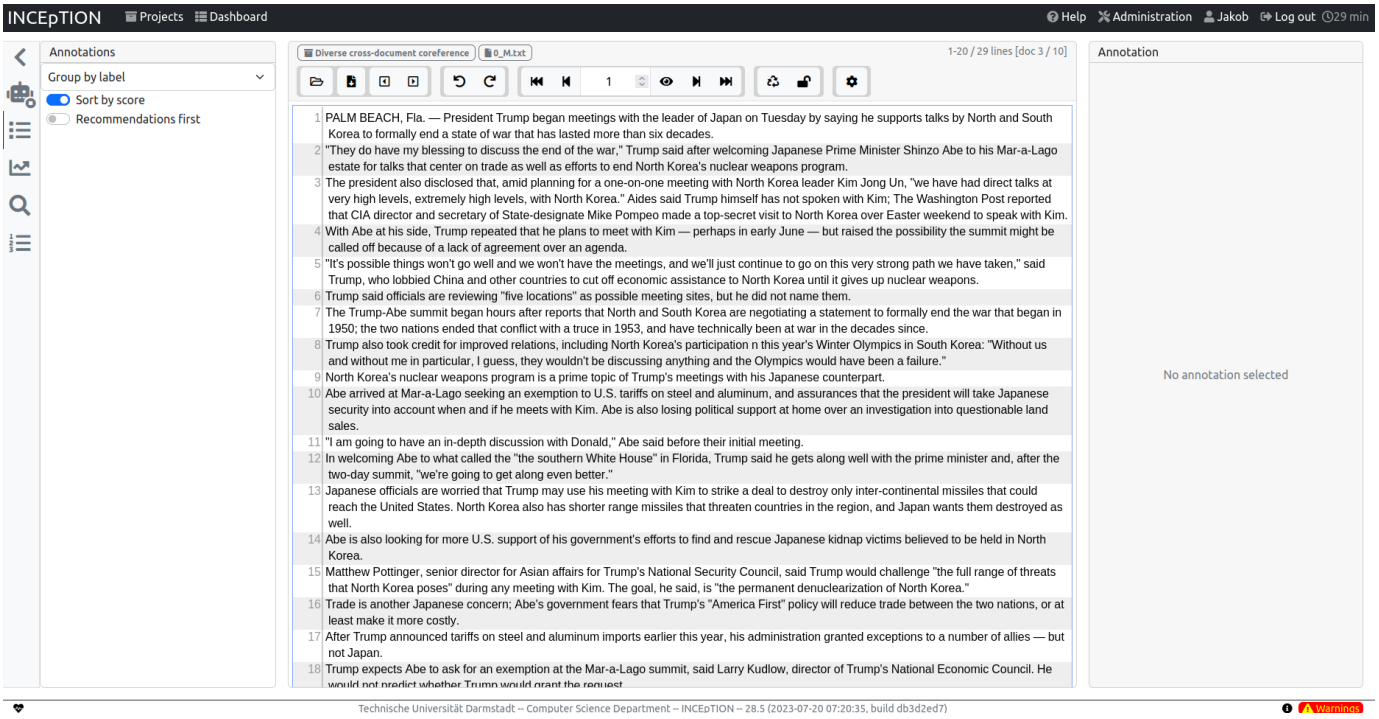

Figure 5: Screenshot of Inception Window Showing a not Yet Annotated Document Loaded into the Annotation GUI.

3.2. User Manual

Inception offers a variety of functionalities of which only those relevant for our project are described here. For a full explanation of how to use Inception, please check the official documentation which can be accessed online or from within the Inception GUI by clicking on ”Help” in the right upper corner. Every annotator’s instance of Inception contains two basic layers of annotation. The first layer, called Entity layer, is triggered when a mention is marked by highlighting text with a simple press-hold-drag mechanism. This opens the layer’s side panel. Here, annotators can fill in the Entity layer’s three parameters:

• **Entity-type:** a drop-down list to select a mention’s entity-type by clicking on or typing the type’s abbreviation.

• **Global Entity-Name:** a mixture of free text field and drop-down list to assign a global entity’s name to a mention. If the name has already been used before, it can be selected as Figure 6: Annotating a mention of ”Donald Trump”: in the right panel, annotators can fill in values for the Entity layer’s three parameters Entity-type, Global entity-name, and Wiki data. Automatically suggested annotations are displayed in Gray boxes above the text rows.

• **Wiki Data:** a search field to type the name of an entity and find its respective Wiki data URI.

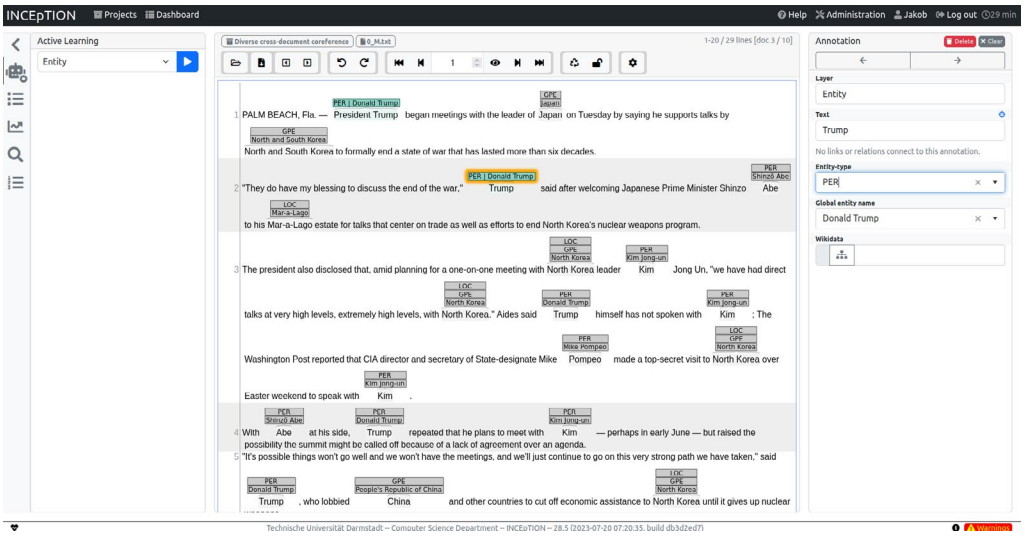

Figure 6: Annotating a Mention of ”Donald Trump”: in the Right Panel, Annotators can fill in Values for the Entity Layer’s Three Parameters Entity-Type, Global Entity-Name, and Wiki data. Automatically Suggested Annotations are Displayed in Gray Boxes Above the Text Rows.

The second layer, called Relation, is triggered when two already marked mentions are connected to each other, again simply by clicking and holding on one mention and dragging the mouse to the other mention. This layer only contains one parameter which is named Label. It is a drop-down list to select a relation-type for labelling the connection between both mentions.

After the first annotations have been made, Inception starts to suggests spans and values for new annotations on the Entity layer. These suggestions are displayed in Gray boxes. One click on a box accepts the suggestion and turns it into a proper annotation, a double-click denies the suggestion and makes the box disappear.

The GUI's upper panel is mostly for navigating through the document. However, it also contains a button for resetting the document by deleting all annotations made so far and a button in the shape of a padlock to mark the annotation process of the document as finished. This button should be pressed at the very end of the annotation, though it is advisable to first annotate each document before marking all of them together as officially finished. Clicking on the gear wheel opens up the GUI's style settings. Here, annotators have the option to adjust panels' margin sizes, the colouring of annotations, and how many text rows are to be displayed simultaneously. Annotations are saved automatically which is why there exists no saving button in the GUI.

Annotators will read each article three times and focus on a different annotation task in each pass: in the first pass, only read the text to get an overview of it. Do not make any annotations, yet. In the second pass, mark mentions with identity-relations, assign an entity to them and link them to Wiki data. In the third pass, annotate near-identity and bridging relations between mentions.

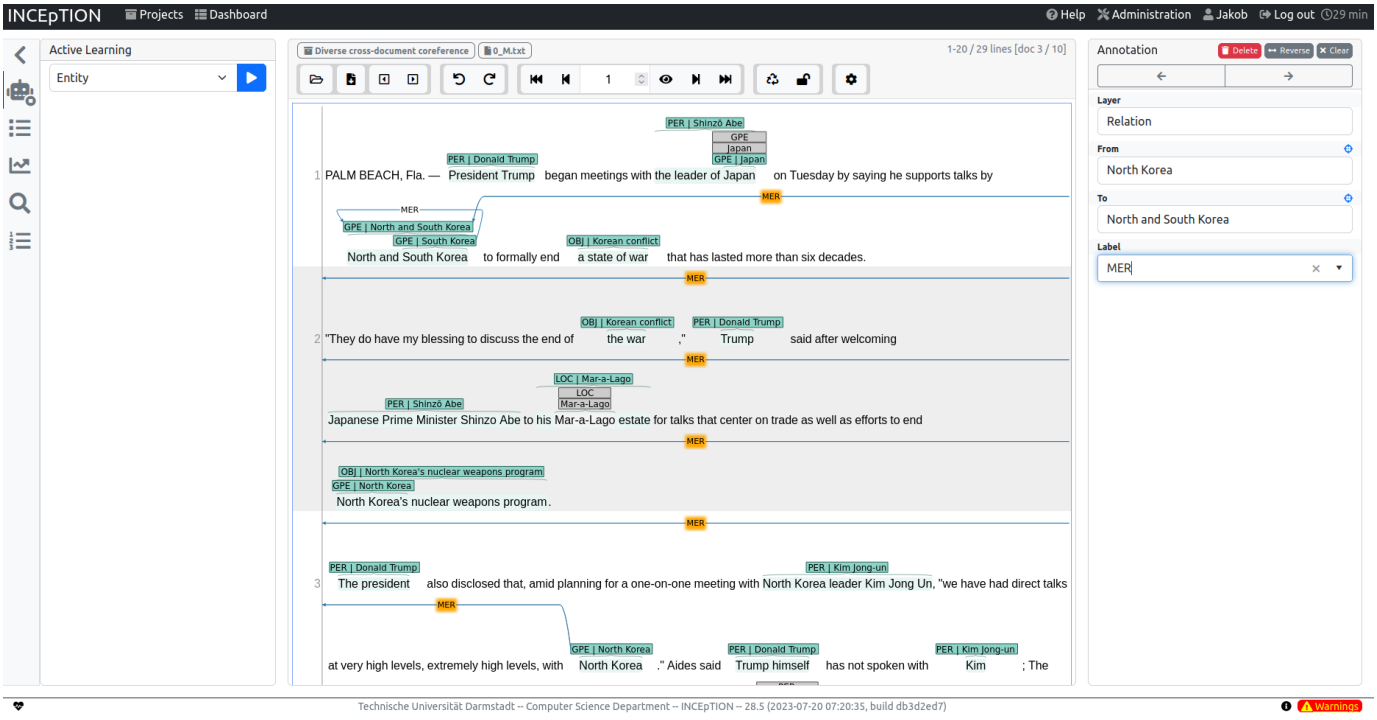

Figure 7: Annotating a Relation Between Two Mentions: the Mention "North Korea" is Connected to "North and South Korea" with a meronymy relation (MER).

4. Annotation Guidelines

Annotators will read each article three times and focus on a different annotation task in each pass: in the first pass, only read the text to get an overview of it. Do not make any annotations, yet. In the second pass, mark mentions with identity-relations, assign an entity to them and link them to Wiki data. In the third pass, annotate near-identity and bridging relations between mentions.

4.1. First Pass: Get Familiar with the Text

Read the entire text carefully. Try to already pay attention to what entities are mentioned, but do not annotate them, yet.

4.2. Second Pass: Annotate Mentions With Identity-Relations

Read the text for a second time. Identify potential coreference candidates. Wherever a referent is referred to by at least two identical mentions, annotate these and all subsequent mentions respectively. Do this as follows:

• First Check if a Candidate is Markable

i In general, only noun phrases (NPs) are markable. This includes nominal phrases ("the president"), proper names ("Mr. Biden"), and quantifier phrases ("all member states").

ii For reasons of efficiency, most pronominal NPs are excluded

from annotation because they normally carry little variation with regards to how they are labelled [6]. However, certain types of pronouns can be included not as head, but as modifier for another NP, e.g. demonstrative pronouns (”this man”) and reflexive pronouns (”the president himself”).

iii Numbers like currency expressions (”€2.3 billion”) and percentages (”19% of the votes”) are included, but dates of any kind (”January 23”, ”1996”, ”this Sunday”) are excluded for now.

iv Given coreferential conjunctions that mention several entities at once and, syntactically, cannot be split (”North and South Korea”), first mark everything that could be extracted as single-entity mention separately (possible for ”South Korea”, but not for ”North”), then mark the entire conjunction. Use a MER-relation to connect mentioned entities with the conjunction (see description of the MER relation in subsection 4.3). • Then check if the candidate you want to annotate is truly identical to other mentions of the same referent. To do so, compare it to the referent’s most previous mention. In case no mention of the referent has been annotated so far, simply compare the two candidates triggering the annotation:

v Identity between two mentions means that both refer to the same entity in almost the same way. In comparison to the first mention, the second one may provide additional information about the referent or only highlight a subset of its attributes, but new and old attributes may not contradict each other [1].

vi When in doubt, ignore all modifiers and focus on the heads of both mentions to check if they are identical.

• If the Candidate is Markable and Identical to Previous Mentions, Start Your Annotation.

First, Mark the Mention:
i. We annotate mentions with a maximum span style. This means that for each candidate, the NP’s head and all of its pre- and post-modifiers are included in the annotation. More precisely, this includes articles (”a”, the”), adjectives (”a worried president”), other NPs (”US president Joe Biden”), appositives (”Joe Biden, president of the United States”), prepositional phrases (”demonstrators in front of the White House”), and relative clauses (”Biden, who was elected president in 2020”) [10]. Any punctuation or white space at the very beginning or end of the span are excluded.

ii. Additionally to maximum span style, we annotate with nested style, meaning a mention’s span may overlap with or contain another mention. But remember not to mark any mention you discover, but only those who actually participate in coreference!

• After Selecting the Correct Span, Assign an Entity-Type to a Marked Mention by Choosing from the Layer’s Respective Drop-Down List

We distinguish between the following entity types: PER, ORG, GRP, GPE, LOC, OBJ.

i. Person (PER): an individual actor.

ii. Organization (ORG): an official organization that is not government-related, e.g. ”the WHO”, ”Fox News”, ”the opposition”.

iii. Group (GRP): a group of individuals acting collectively or sharing the same properties, e.g. ”demonstrators”, ”unemployed beneficiaries”, ”the two leaders”.

iv. Geo-Political Entity (GPE): a state, country, province etc. that comprises a government, a population, a physical location, and a nation [11]. This includes clusters of GPEs, e.g. ”Eastern Europe” or ”the Arab League”. Governmental organizations or locations that represent an entire GPE are also marked as GPE, e.g. ”the US government”, ”US officials”, ”the Biden administration”, ”Washington”, ”the White House”.

v. Location (LOC): A physical location that is not a GPE, e.g. ”Los Angeles”. This includes mentions like ”Germany” or ”the White House” when referred to not in a political way, but with a focus on its geographic, cultural, architectural and other locality attributes. Be aware that two mentions with the same textual representation but different entity-types are not to be marked as identical! Instead, most of such cases would imply a MET-relation.

vi. Object (OBJ): An object or other concept that is mentioned, e.g. ”Biden’s hands”, ”a submarine”, ”the results”. However, objects are static concepts. Do not confuse them with NPs that express events or other changes of state (”election”, ”negotiations”, ”Biden’s statement”) which we do not annotate! • Now it is time to assign the mention to an entity cluster. With this step, you create or extend a local coreference chain. At the same time, you link it with corresponding discourse entities across documents and globally with its actual referent.

vii. In case that, in the present document, you already have annotated previous mentions of the same entity, you will also already have created a local coreference cluster. The cluster will already be linked to a global discourse entity and to a referent. To assign the current mention to that cluster, select the global entity’s name from the respective drop-down list. The Wiki data field can be left empty.

viii. If, on the other hand, no previous mentions have been annotated, you are faced with two identical mentions you want to create a new local cluster of. To do this, first fill in the fields of the first mention. ∗ Begin with the Wiki data field and type in the referent’s name. Inception now looks for a suiting Wiki data entry and displays a drop-down list with the search results. Select the correct entry from that list. To enhance search results, try to look for the entity’s most neutral name, ignoring articles. Sometimes it is easier to look for the entry on the Wiki data website itself and then copy its name into the field. If no Wiki data entry exists, leave the field empty. ∗ Assuming you have

found a Wiki data entry, copy the text displayed in the Wiki data field into the Global entity name field. By doing this, the name will automatically be added to the underlying tag set, meaning you will be able to select it from the drop-down list in subsequent annotations. However, if you have not found a Wiki data entry, copy the mention's text, again with maximum span style, into the Global entity-name field. Use this text as name for any following coreferential mentions. If the name has already been used for a semantically different entity in another document, add the document ID to the new name.

ix. Now turn to the second mention and annotate it based on the previous one. That is, assign the Global entity-name while leaving the Wiki data field empty.

4.3. Third Pass: Annotate Mentions with Different Relations

Read the text for a third time. Wherever you see two mentions connected through a near-identity relation, make a respective annotation: • For every new mention that has not been marked in the second pass already, check if it is markable and annotate it with its correct span and entity-type as described above. However, leave the Global entity-name and Wiki data field empty. • When both mentions are marked with the correct span and entity-type, connect them with one of the following near-identity relation types: MET, MER, CLS, STF, DEC, BRD [1,12-14].

a. Metonymy (MET): In a MET-relation, in comparison to its antecedent, an anaphor highlights different facets of an entity. This includes facets like: ∗ a certain role or function performed by an entity. Consider example (5).

(5) "Although *Biden is head of the Democrats, he is also president of all Americans."*
Assuming "*Biden*" has already been annotated as part of a respective cluster in the second pass, *"head of the Democrats"* and *"president of all Americans"* would now be connected to "*Biden*" with a MET-relation. However, in this example, it is the juxtaposition of both roles in particular that makes this a case of metonymy. In a more regular context, naming one of these roles alone could be annotated in the second pass as identical mention, instead.

∗ a location's name to refer to an associated entity, e.g. "*Washington*" as metonym for *"the US government", "China"* for *"the Chinese government", "Silicon Valley"* for *"the Tech industry".*

semantically different group of people. In this case, do not change your annotations of the previous document, but do use the Global entity-name "*demonstrators0 L*" in the current document.

∗ An organization's name to refer to an associated place, e.g. a bank's name like "*ECB*" to refer to the building that contains that bank's headquarters.
∗ different forms of realization of the same piece of information, like in example (6), where the same content is manifested once as audible speech and once as written text.

(6) "Though it is questionable whether he had actually written *the piece* himself, Macron gave *a truly brilliant speech* this afternoon."

∗ Representation, where one mention is a picture or other representation of an entity, as already seen in example (4).

(4) "The AfD is circulating *a photo of Angela Merkel with a Hijab,* although *Merkel* never wore Muslim clothes." ∗ other facets, since this is no exhaustive list and metonymy is a dynamic phenomenon.

∗ given two ID-clusters that are metonymous to each other (e.g. several mentions of "the US president" and several mentions of "the White House" which often participate in metonymy together), do not connect every single mention of the latter to a mention of the former, but only do this for the latter's first truly coreferential mention.

b. Meronymy (MER): A MER-relation between two mentions indicates that: ∗ one mention is a constituent part of the other in whatever direction, as in example (7).

(7) *"President Biden* expressed his concern about the ongoing ... *'The US government* will not ...', he stated." ∗ one mention refers to an object which is made of the stuff which the other mention refers to.

(8) "The duty on *tobacco* has risen once again, making *cigarettes* as expensive as never before."

∗ both mentions refer to overlapping sets.
(9) *"AfD supporters* demonstrated in front of the Reichstag this morning. Among *the crowd* was ..."

∗ Finally, a MER-relation can be used to specify entities mentioned in syntactically non-dividable conjunctions. Given such a conjunction, as "North and South Korea" in example (10), mark "South Korea" separately as it can be treated as independent noun phrase. The adjective phrase "North", however, cannot be marked. Instead, mark the entire conjunction and connect "South Korea" to it with a MER-relation (illustrated by the dotted underlining). Do the same for the first full mention of "North Korea" that follows in the text. If none follows, use a previous mention or, if there is none, ignore the "North" mention.

(10) *"North and South.....Korea* have resumed negotiations ... *North Korea seems ...*"
– **Class** (CLS): a CLS-relation indicates an 'is-a' connection between two mentions. One mention thus belongs to a sub- or superclass of another.
–
(11) "In way, *Trump* only seized the opportunity. This is what *skilled politicians* do."

c. Spatio-Temporal Function (STF): a mention refers to an entity that deviates in place, time (3), number, or person (12).

(3) ”Even if *the young Erdogan* used to be pro-Western, *Turkey's president* nowadays often acts against Western interests.” (12) ”A historic meeting: *a pope* and *a pope* shaking hands.”

d. Declarative (DEC): where two mentions X and Y are connected through verbal phrases like ”X seems like Y”, ”stated that X was Y”, ”declared X Y”, or other declarations as in (13), they can be connected with a DEC-relation.

(13) ”In his speech, he also spoke about *North Korea* and called it a *fundamentally barbaric nation.*”

The DEC-relation thus includes definitions and descriptions of entities. This is especially the case when declarative clauses are used within quotes. However, when value-free declarative clauses like ”X is Y” are used as quasi objective specifications of an entity, they might indicate an identity relation, instead. The same structure might be used to assign a super-class to the entity, making it a CLS-relation.

e. Bridging (BRD): for reasons of simplicity, we have included BRD in our subsumption of different relation-types under the term of near-identy. Despite of that, BRD is actually a separate phenomenon from both identity and near identity. BRD connects two entities that are mostly independent of each other while nonetheless, the existence of one can be inferred by the existence of the other [15]. Technically, the BRD-relation could be used to mark all sorts of ontological connections between entities. This is not the purpose of this annotation scheme, though. Instead, we use BRD only where the mention of one entity influences the depiction of an associated entity or where one entity is modified by a possessive pronoun that refers to another entity. Example

(14) Illustrates both use cases:
(14) ”Unlike *Queen Elizabeth, Charles* has not been shy about promoting *his political views.*”

Here, the NP *”his political views”* contains a modifying possessive pronoun, which is why it is to be annotated as bridging to ”*Charles*”. Additionally, the mention ”*Charles*” can only be interpreted correctly as referring to Charles III (and not any other Charles) by its juxtaposition with the NP ”*Queen Elizabeth*”. Hence ”*Charles*” is to be annotated as bridging to ”*Queen Elizabeth*”. • Deciding on what relation-type to choose can be difficult. When in doubt, follow these general guidelines:
– use an identity relation rather than a near-identity relation (especially DEC).
– when having to choose between near identity relations, use MET rather than MER.
– use MER rather than CLS.
– use CLS rather than DEC.
– use any near-identity relation that is not BRD rather than BRD.
• When annotating near-identity and bridging, always connect an anaphoric mention to the nearest possible antecedent. But remember that antecedents normally appear before an anaphor. Only if necessary you may connect a mention to a subsequent expression (making their relation cataphoric).

5. Conclusion and Future Work

Our proposed annotation scheme covers a multitude of coreferential relations. It gives a detailed explanation of how to mark coreferential mentions across documents, assign entity-types and names to them, connect them with each other, and link them to the Wiki data knowledge graph. The scheme thus represents a significant step toward more accurately capturing the complexities of coreference use. It furthermore provides a valuable resource for researchers both in the field of coreference resolution and media bias by word-choice and labelling. Having said that, our scheme leaves room for possible extensions to further advance research in those domains. First, the annotation of events could be included in our scheme. An interesting question that arises is whether the relation types as outlined here could be applied not only to entities, but to events all the same. A second possible extension would be to include a layer of media bias annotation to the scheme, enabling a direct comparison of diverse coreference usage and media bias by word-choice and labelling. Both proposed extensions could be easily added on top of our scheme. Having said that, the present form of our scheme already addresses many of the complexities of diverse cross-document coreference and offers a roadmap for capturing nuanced linguistic relationships, ultimately advancing our understanding of language and discourse in digital print media.

Acknowledgements

Many thanks to my project supervisor Anastasia Zhukova who never became tired of my many questions and always knew how to help me out with good advice whenever I felt stuck.

References

1. Recasens, M., Hovy, E. H., & Martí, M. A. (2010, May). A Typology of Near-Identity Relations for Coreference (NIDENT). In LREC.
2. Hamborg, F., Zhukova, A., & Gipp, B. (2019, June). Automated identification of media bias by word choice and labeling in news articles. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 196-205). IEEE.
3. Royce Kurmelovs. 2023. Fox News labels Joe Biden a 'wannabe dictator' during Trump speech. The Guardian. https://www.theguardian.com/media/2023/jun/14/foxnews-labels-joe-biden-a-wannabe-dictatorduring- trump-speech, accessed 13 Sep 2023
4. Michael Luciano. 2023. Trump Debuts New Nickname for Biden at New Hampshire Rally. Mediaite. https://www.mediaite.com/politics/trumpdebuts-new-nickname-for-biden-at-newhampshire-rally/, accessed 13 Sep 2023.
5. Liu, R., Mao, R., Luu, A. T., & Cambria, E. (2023). A brief survey on recent advances in coreference resolution. Artificial Intelligence Review, 1-43.
6. Zhukova, A., Hamborg, F., Donnay, K., & Gipp, B. (2022, February). XCoref: Cross-document coreference resolution in the wild. In International Conference on Information (pp. 272-291). Cham: Springer International Publishing.
7. Spinde, T., Kreuter, C., Gaissmaier, W., Hamborg, F., Gipp, B., & Giese, H. (2021, September). Do you think it's biased? how to ask for the perception of media bias. In 2021 ACM/

IEEE Joint Conference on Digital Libraries (JCDL) (pp. 61-69). IEEE.
8. Spinde, T., Krieger, D., Plank, M., & Gipp, B. (2021, September). Towards a reliable ground-truth for biased language detection. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 324-325). IEEE.
9. Klie, J. C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018, August). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In proceedings of the 27th international conference on computational linguistics: system demonstrations (pp. 5-9).
10. Hirschman, L., & Chinchor, N. (1998). Appendix F: MUC-7 coreference task definition (version 3.0). In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.
11. Linguistic Data Consortium. 2008. ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf, accessed 18 Sep 2023.
12. Clark, H. H., & Bangerter, A. (2004). Changing ideas about reference. In Experimental pragmatics (pp. 25-49). London: Palgrave Macmillan UK..
13. Nedoluzhko, A., Mírovský, J., & Pajas, P. (2009, August). The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague Dependency Treebank. In Proceedings of the Third Linguistic Annotation Workshop (LAW III) (pp. 108-111).
14. Spala, S., Miller, N. A., Yang, Y., Dernoncourt, F., & Dockhorn, C. (2019, August). DEFT: A corpus for definition extraction in free-and semi-structured text. In Proceedings of the 13th Linguistic Annotation Workshop (pp. 124-131).
15. Clark, H., & Haviland, S. E. (1977). Comprehension and the given-new contract. Discourse production and comprehension, ed. by Roy O. Freedle, 1-40. Hillsdale, NJ: Erlbaum, 1, 40.
16. Ad Fontes Media, Inc. 2023. Media Bias Chart, Version 2.8.4. https://adfontesmedia.com/interactive-mediabias-chart/, accessed 06 Jul 2023.
17. AllSides Technologies Inc. 2023. AllSides Media Bias Chart, Version 9.0. https://www.allsides.com/media-bias/mediabias-chart, accessed 06 Jul 2023
18. O'Gorman, T., Wright-Bettner, K., & Palmer, M. (2016, November). Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In Proceedings of the 2nd workshop on computing news storylines (CNS 2016) (pp. 47-56).
19. Zhukova, A., Hamborg, F., Donnay, K., & Gipp, B. (2022, February). XCoref: Cross-document coreference resolution in the wild. In International Conference on Information (pp. 272-291). Cham: Springer International Publishing.
20. Linguistic Data Consortium. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events, Version 5.4.3. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf, accessed 22 Sep 2023.
21. Michael Luciano. 2023. Trump Debuts New Nickname for Biden at New Hampshire Rally. Mediaite. https://www.mediaite.com/politics/trumpdebuts- new-nickname-for-biden-at-newhampshire- rally/, accessed 13 Sep 2023.