A checklist for coreference resolution papers and what it tells us about reproducibility in natural language processing

K. Bretonnel Cohen [1,*], You [2] and Lawrence E. Hunter [1]

[1] *Biomedical Text Mining Group, Department of Biomedical Informatics, U. Colorado School of Medicine, Aurora, Colorado, USA*
[2] *Laboratory X, Institute X, Department X, Organization X, City X , State XX (only USA, Canada and Australia), Country X*

Correspondence*:
Corresponding Author
kevin.cohen@gmail.com

ABSTRACT

There is currently a lot of variation in how papers about coreference resolution report their materials, methods, and results. Most papers on coreference resolution lack so much important information that it is difficult to interpret their findings. For example, 9/17 papers on coreference resolution do not give the number of markables in their data, and 8/17 papers on coreference resolution do not give the distribution of markables. We present a checklist and Minimum Information reporting standard that is field-tested and can help to design research projects, write papers, and improve reproducibility of research on coreference resolution.

Keywords: reproducibility, coreference resolution, anaphora resolution, natural language processing, biomedical natural language processing, text mining, checklists, Minimum Information standards

1 INTRODUCTION

It has become difficult to ignore: an increasingly popular research and clinical tool—natural language processing—is plagued by problems of methodology (Fokkens et al., 2013; Crane, 2018; Mieskes et al., 2019a,b), reporting practices (Cohen et al., 2017a; Mieskes, 2017), and publishing practices (Pedersen, 2008; Cohen et al., 2020; Sedoc et al., 2021; Bianchi and Hovy, 2021; Sedoc et al., 2021). All of these issues have been identified as proximate causes of a ubiquitous problem in scientific research, known today as the Reproducibility Crisis. The reproducibility problem is a difficult one—so difficult that although the authors of this paper have studied it for years from many different perspectives (e.g. Caporaso et al. (2008); Johnson et al. (2007); Branco et al. (2017); Cohen et al. (2020); Vos et al. (2020)), we still have relevant reproducibility failures of our own Cohen et al. (2016b, 2018). In this article, we propose, develop, and evaluate a tool for addressing reproducibility in natural language processing via the issues of methodology, reporting practices, and publishing practices: a reproducibility checklist.

Checklists are demonstrably useful for many tasks, both high-level and low-level[1]. It is likely that checklists could help us address the many forms of reproducibility problems that are mentioned in the preceding paragraph. But, designing a *good* checklist requires some science, some human factors engineering, expert input, non-expert testing, and a lot of experimentation. Consequently, there are not nearly enough checklists for the vast majority of scientific fields, and the number for natural language processing is actually quite small—nor, in general, have the existing ones actually been evaluated, empirically or otherwise.

1.1 What is relevant to reproducibility?

You would think that computational experiments would be easy to replicate. After all, they *can* be done with no manual intervention; all code and data *can* be stored; and computation *can* be deterministic. Nonetheless, a wide array of factors have been found to hinder both the replicability of computational experiments and the reproducibility of their findings. Not everything in the universe—yet—but, probably more than you would think. Fokkens et al. give a nice sketch of some unexpected things that affected reproducibility of results in *one set* of experiments—treatment of ties, rounding of real numbers, different splits of the training/testing data, versions of resources such as WordNet…Fokkens et al. (2013). Unexpected things, and typically undocumented things—but not *unanticipatable* things, and hence the value of reproducibility checklists.

1.2 What is coreference resolution?

Coreference is the linguistic phenomenon of having multiple things in a discourse (e.g. a written text, a conversation, or a tweet) that refer to the same concept or thing in the world. *Coreference resolution* is a computational task that can be defined as follows: given an input that makes multiple reference to concepts or things in the world, identify the sets of references in a specific text to the same concepts or things. Besides its high relevance to linguistic and psychological theory McCawley (1983, 1976); Reinhart (1983); McCawley (1998); Tavares et al. (2015); Wongkoblap et al. (2021), as well as to clinical neurology Alves et al. (2021), it is a technical challenge in an increasingly common biomedical research tool known as natural language processing (the use of computers to analyze linguistic data such as scientific journal articles and electronic health records). See Choi et al. (2014); Gîfu and Iliescu (2015); Choi et al. (2016b); Gîfu and Cioca (2017); Gîfu and Onofrei (2017); Wongkoblap et al. (2021) for illustrations of some of the linguistic and semantic phenomena that make it difficult.

1.3 What is (fairly) unique to coreference resolution?

Perhaps you ask if we need yet another natural language processing reproducibility checklist, specificially for coreference resolution? The question itself is not well-formed, because we do not yet have *any* community-consensus reproducibility checklists for natural language processing. But, what if we did? Would we still need one for coreference resolution? Yes, because the data and performance metrics that are used in coreference resolution research are special.

The evaluation data that is used in coreference research has one characteristic that other natural language processing datasets do not: the notion of the *markable*. Think of the markable as the minimal unit of analysis in coreference resolution: the smallest unit that could potentially take part in a coreferential

[1] For example, the act of preparing a good evaluation checklist helps us to answer the very high-level question "how should this kind of research be evaluated" **?**? Using a checklist also helps us with answering low-level questions like "Can I actually hit the "Submit" button for my ACL paper, or not?" Some notable examples of the use of checklists include for handling in-flight aircraft emergencies, for medical procedures, and for maintaining tanks Dunnigan (2003).

or anaphoric relation. The earliest definition of *markable* that we have found is that of Hirschman and Chinchor Hirschman and Chinchor (1998)[2]. On that definition, a markable is any noun, noun phrase, or pronoun. Later work has sometimes advocated for a more restrictive definition, but this one works quite well. In fact, more restrictive definitions have led to problems of reproducibility in natural language processing research. How so?

Being unique to phenomena of coreference and anaphoric reference, markables are only represented in coreference/anaphora resolution datasets. In those datasets, it is crucial to know (a) how many there are, and (b) what their distribution is. If we do not, we cannot really evaluate the research. Overly restrictive definitions of "markable" reduce the number of annotations in a gold standard, and can completely eliminate even the *possibility* of detecting at least two classes of system errors. For example: some markables are *potentially* coreferential with others, but in fact are not. They are known as *singletons*. It turns out that some datasets represent singletons, while some ignore them entirely; some research papers use datasets in which singletons are marked, but do not take them into account in calculating performance measures. Inclusion or exclusion of singletons can have an enormous effect on performance measures Kübler and Zhekova (2011); Recasens et al. (2013); Pradhan et al. (2014). The magnitude of that effect is proportional to the number of singletons in the data. So, this element of coreference resolution research papers is both crucial *and* unique to coreference resolution. For another example: some datasets mark non-referential pronouns (e.g. *they* and *it* in *they say it's going to rain*.). Others do not. The distinction has historically had many implications for theoretical linguistics Peled (1990); Zimmermann (2014). Distinguishing between referential and non-referential pronouns is itself a non-trivial task in natural language processing Bergsma et al. (2008), so knowing whether or not a gold standard evaluation set expects systems to make that distinction is important to evaluating any performance claims.

Our conclusion: the topic needs its own reproducibility checklist.

2 MATERIALS AND METHODS

It is certainly possible to make a *bad* checklist. Tables **??** and **??** give examples of bad checklists for coreference resolution systems. So, to prepare *this* checklist, we began with the methods of the Evaluation Checklist Project. Those methods have many similarities to a modified Delphi method.

1. The Evaluation Checklist Project focuses in its initial steps on the importance of expert input. So, two experienced coreference resolution researchers (authors KBC and MP) each constructed an initial draft of the checklist.
2. The next step of the Evaluation Checklist Project's procedure is to prepare a draft for field testing. To do this, we discussed and harmonized the two individually-created drafts, and from that, we assembled the next draft.
3. We distributed the field-test draft via...

We did two field tests of the checklist:

1. We analyzed a set of published papers on coreference resolution, looking for documentation of selected items in the checklist. Specifically, we looked for two things that we identify above as specific to coreference resolution research: (1) the count of markables, and (2) the distribution of markables.

[2] There may be an earlier one from MUC-6, but we have not found one.

2. We used the checklist to design and write a paper about our own ongoing project on coreference resolution.

For field testing, we used the following published papers: Zhekova and Kübler (2013); Lee et al. (2017); Joshi et al. (2019); Kantor and Globerson (2019); Cao and Daumé III (2020); Uppunda et al. (2021); Yin et al. (2021); Rudinger et al. (2018); Poesio et al. (2004); Chen et al. (2011); Cybulska and Vossen (2015); Jauhar et al. (2015); Aktaş et al. (2020); Lapshinova-Koltunski et al. (2020); Wilkens et al. (2020); Aloraini and Poesio (2021); Yu et al. (2021) These seventeen papers, years of publication from 2004–2022 **as of 2022-12-22** were reviewed for *Markables — count* and *Markables — distributional analysis of*. We focussed on these items because they are the most unique to coreference resolution (see Section 1.3 above).

3 RESULTS

3.1 Field testing

8/17 gave a count of markables. 9/17 gave a distributional analysis of markables[3]. For something that is as crucial to evaluating coreference resolution research, this is a very low number.

The ECP criteria for evaluating an evaluation checklist are organized into six categories:

1. Appropriateness of content
2. Clarity of purpose
3. Completeness and relevance
4. Organization
5. Clarity of writing
6. References and sources

Each of these categories contains two or more items, of varying degrees of granularity. In the subsequent sections, we discuss our proposed checklist in terms of those items. Most of them were clearly applicable to this checklist; the one exception is noted.

3.2 Appropriateness of content

Quoted directly from the ECP[4]:

- *The checklist addresses one or more specific evaluation tasks (e.g., a discrete task or an activity that cuts across multiple tasks).*
- *The checklist clarifies or simplifies complex content to guide the performance of evaluation tasks.*
- *Content is based on credible sources, including the author's experience.*
- *Content is consistent with the program evaluation standards (Yarbrough, Shulha, Hopson, and Caruthers, 2011) and the American Evaluation Association's Guiding Principles for Evaluators (2013) and Statement on Cultural Competence in Evaluation (2011).*
- *Content does not overtly favor one evaluation approach over others unless the checklist is intended to support the application of a particular evaluation approach.*

Here is how we addressed each of those checklist items:

[3] We find it surprising that *any* paper would give a distributional analysis without also giving counts. But, one did.

[4] https://wmich.edu/evaluation/checklists/checklistsvalidation

T"he checklist clarifies or simplifies complex content to guide the performance of evaluation tasks." (Direct quote from ECP) We clarified how to record complex content by suggesting a format for each item on the list. For example, the items *Figures of merit* and *Parameters* should have lists as their values. In contrast, the *Algorithm category* item would be expected to be *one* of *Rule-based, Machine learning (supervised), Machine learning (unsupervised),* or *Sieve*.

"Content is based on credible sources, including the author's experience." Our use of credible sources is reflected in the size of the bibliography. With respect to the authors' experience: MP is a world-class expert in the field of coreference resolution. KBC wrote the earliest paper on reproducibility in natural language processing, to the best of our knowledge.

"Content is consistent with the program evaluation standards (Yarbrough, Shulha, Hopson, and Caruthers, 2011) and the American Evaluation Association's Guiding Principles for Evaluators (2013) and Statement on Cultural Competence in Evaluation (2011)." To address these items, we worked off of the summary page of the Statement on Cultural Competence in Evaluation, because the link to the full statement is broken. Following its guidance, we maximized the diversity of AEA-sanctioned cultural factors, including the following in the list of authors:

1. One upper-class author, one trailer trash author
2. One white person, one Jew
3. One native speaker of English, one native speaker of Italian, and one member of two communities suffering linguistic oppression
4. One cis-gender straight person, one LGBTQQIP2SAA person
5. One North American, one Southern European
6. We were not able to assure diversity of lineage because we do not know what *lineage* is. We think that it might have something to do with nobility (inherited aristocracy) and royalty (the rank above nobility). We erred on the side of including only commoners in the author list.
7. No members of the author list belong to a caste.

"Content does not overtly favor one evaluation approach over others unless the checklist is intended to support the application of a particular evaluation approach." The checklist is intended to support—and therefore favors—evaluation in terms of a normative conception of traditional Western European logic Steinberger (2022) and a falsificationist philosophy of science (Popper (1953), and also see Gordin (2012) for an excellent treatment of related issues). As such, it does not support a role for intuition, dialectical logic, or non-binary logics. (See Chapter 7 of Okasha (2002) for perspective.) This is obviously a limitation of the work.

3.3 Clarity of purpose

- *A succinct title clearly identifies what the checklist is about.*
- *A brief introduction orients the user to the checklist's purpose, including the following:*
  - *The circumstances in which it should be used*
  - *How it should be used (including caveats about how it should not be used if needed)*
  - *Intended users*

Here is how we addressed each of those checklist items:

"A succinct title clearly identifies what the checklist is about." The checklist works as a checklist and as a Minimum Information standard, and it is intended for use with coreference resolution research and development, so we titled it *A checklist and Minimum Information standard for corefereference resolution research and development*.

"A brief introduction orients the user to the checklist's purpose, including… [t]he circumstances in which it should be used." The checklist is primarily intended for use while *planning* and *doing* research and development. Evaluation of submitted research papers by editors and peer reviewers is a secondary intended use.

"A brief introduction orients the user to the checklist's purpose, including… [h]ow it should be used." The checklist should be used while planning research, as a reminder of some of the many variables that can affect the interpretability and the likely generalizability (or lack thereof) of research results and of commercial products. It should be used while doing research to record the many system issues that are crucial to the ability to repeat experimental methodologies and to interpret research results.

"A brief introduction orients the user to the checklist's purpose, including… [i]ntended users." The checklist is intended for use by researchers and by developers, and secondarily intended for use by editors and by peer reviewers.

3.4 References and sources

Direct quote from the EPC:

- *Sources used to develop the checklist's content are cited.*
- *Additional resources are listed for users who wish to learn more about the topic.*
- *A preferred citation for the checklist is included (at the end or beginning of the checklist).*
- *The author's contact information is included.*

Here is how we addressed each of those checklist items:

"Sources used to develop the checklist's content are cited." The sources are included in this paper, including work on checklists, on coreference resolution, and on reproducibility.

"Additional resources are listed for users who wish to learn more about the topic."

- Systematic review in English: Uzuner et al. (2012)
- Review articles in English: Zheng et al. (2011); Sukthanker et al. (2020); Olex and McInnes (2021)
- Book chapter in English: McShane and Nirenburg (2021)
- Book in English: Mitkov (2014)
- Books in French: Poibeau (2003, 2011)

In Chinese: Lang et al. (2007); and (2015); et al. (2019)

Examples of domain-specific and task-specific applications of coreference resolution are useful for understanding how coreference resolution interacts with linguistic and structural particularities of your data. See, for example, Apostolova and Demner-Fushman (2009); Grouin et al. (2011); Apostolova et al. (2012); Bodnari et al. (2012); Kim et al. (2012); Nguyen et al. (2012); Uzuner et al. (2012); Chowdhury and Zweigenbaum (2013); Kilicoglu and Demner-Fushman (2014); Choi et al. (2015); Lavergne et al. (2015); Choi et al. (2016a); Kilicoglu and Demner-Fushman (2016); Fang et al. (2022).

"A preferred citation for the checklist is included (at the end or beginning of the checklist)." A bibliographic entry for this paper (the preferred citation) is included at the beginning of the checklist.

"The author's contact information is included." Including an author's contact information is more complicated than one might think, and reproducibility experiments have stumbled on this very issue. For example, the first author of this paper has included his contact information, but due to the brutal nature of the Russian invasion of Ukraine, it is not clear how much longer he will be alive to check his email Plokhy (2015); Applebaum (2018); Plokhy (2018); Chhugani et al. (2022); Catoire (2022); Zhang et al. (2022). We have also included the senior author's contact information, but his email address is an institutional one, so it will eventually stop working, if only due to his eventual retirement. We also note in passing the problem of fake *author* (not reviewer) email addresses in scientific publications Gu J and Z (2015); Dyer (2016); Wang et al. (2022)[5].

[5] See the Retraction Watch web site for details on the author email falsifications involved in these retractions.

Table 1. A bad reproducibility checklist for coreference resolution. It violates all three principles of checklist organization: (1) logical ordering, (2) division into sections, and (3) breakdown of complex components (e.g. "Algorithm" is one single component that should be several).

1	Algorithm
2	APPOS chain count
3	Baseline
4	Code location
5	Configuration parameters
6	Data location
7	Error analysis categories
8	Experimental parameters
9	Figures of merit
10	IDENT chain count
11	Knowledge sources
12	Location of intermediate outputs
13	Markable count
14	Markable distribution
15	Named entity count
16	Named entity taggers
17	Paragraph splitter
18	Parser
19	POS tagger
20	Rule types
21	Rules
22	Semantic class count
23	Sentence count
24	Sentence splitter
25	Source of data for tables and figures
26	Stemmer
27	Token count
28	Tokenizer
29	Word count

Table 2. A bad reproducibility checklist for coreference resolution. It violates the clarity and complexity principles of checklist design. For example, *Materials* is undefined (does it include only the test data, or web resources and dictionary versions, too?); what are the relationships and differences between *Evaluation* and *Results?*

1	Method
2	Materials
3	Evaluation
4	Results
5	Availability

3.5 Limitations and questions for future work

This paper leaves some questions unexplored, and they present fruitful opportunities for future research.

- We field-tested the checklist only on methodology development papers. It should be tested for generalizability on papers that present new coreference datasets, e.g. Lapshinova-Koltunski et al. (2022), which is freely available on the PapersWithCode web site. They can be expected to present some new issues, e.g. inter-annotator agreement and sampling of/exclusion criteria for texts.

Table 3. A bad reproducibility checklist for coreference resolution. It violates the principle of clarity of purpose: is this meant for coreference resolution, or is it limited to pronominal anaphora resolution?

1	Method
2	Materials
2a	Pronoun distribution: referential and non-referential
2b	Pronoun distribution: singular versus plural
3	Results
4	Availability

- More papers on shared tasks and on participation in shared tasks should be included. Some shared tasks have required reporting guidelines, and these could be good tests for the general topic of reproducibility checklists.
- We did not stratify sampling across publication venues or domains. This is important because if there are marked differences between, say, PubMed-indexed papers and *ACLverse papers, then one community might be able to learn a lot from the other. Also, the clinical biomedical domain often has very different data privacy issues from other domains, and this introduces some challenges to the documentation of distributional characteristics in the data.
- The paper contains no real analysis of changes in reporting practices over time. This is relevant because if the field is improving—however unlikely that might be at the time of publication—then maybe it is not as urgent as we suspect to implement this paper's suggestions. On the other hand, if it is *not* improving, then that would lend some urgency to the paper's thesis.
- More field testing is rarely a bad idea. Here we limited ourselves to the three-person maximum that is recommended for usability testing, but that limits diversity of all kinds in the evaluation population, while some kinds of diversity are almost certainly experimentally relevant—for example, different levels of experience (students versus post-doctoral fellows versus senior researchers), resource-rich research environments versus resource-poor environments…

4 CONCLUSIONS

There is currently a lot of variation in how papers about coreference resolution report their materials, methods, and results. Most papers on coreference resolution lack so much important information that it is difficult to interpret their findings. For example, 9/17 papers on coreference resolution do not give the number of markables in their data, and 8/17 papers on coreference resolution do not give the distribution of markables. This paper presents a checklist and Minimum Information reporting standard that is field-tested and can help to design research projects, write papers, and improve reproducibility of research on coreference resolution. Our literature review suggests that the topic is novel[6], and that it should be of interest to many researchers: there are many papers that mention both *coreference* and *reproducibility*, without giving a clear approach to enhancing reproducibility in coreference resolution research. To illustrate this point, Table **??** lists all of the papers on the first page of the Google Scholar search results for the query `reproducibility coreference resolution`. We are gratified to note that one of them actually does present a reproducibility checklist **?**. The rest (including one of our own papers) mention reproducibility either notionally, or simply in passing.

[6] allintitle: reproducibility coreference — zero results.

Table 4. The large number of papers that mention both coreference and reproducibility suggests that this work would be of interest to many researchers. This table shows the first page of results from the Google Scholar search *reproducibility coreference resolution*.

We know that it is not necessarily easy to utilize any method for structuring information for the first time. Indeed, we do not need to look any further than the review of related work in one of our own papers on coreference, Cohen et al. (2017b), to see this. In that paper we tried to give the sizes and other quantitative descriptors for all previously published biomedical coreference corpora. We found that there was so much diversity in how different papers described the size of the associated corpus that all four of the previously published corpora had differently structured quantitative descriptions (see Tables 1, 2, 5, and 6 of Cohen et al. (2017b)).

But, that does not mean that it is not doable, and it *certainly* does not mean that it is not worth becoming comfortable with the process. It is difficult to believe that so much effort would have been put into developing informatics checklists if it were not worth the trouble.

ETHICAL ISSUES CONSIDERED

Equity of pay for annotators has long been known to be a source of ethical problems in natural language processing Fort et al. (2011); Cohen et al. (2016a). At the time of writing, one of the annotators has not been paid at all for two half-months of work.

In the course of writing this paper, we took the opportunity to evaluate ChatGPT's ability to do the writing. It violated rules of professional conduct, fabricating citations and failing to cite relevant work (see the supplementary materials on GitHub) Ray et al. (2022).

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Author contributions are specified in Table 5. We followed the Committee on Publication Ethics guidelines[7] in determining inclusion in the list of authors (see Wager (2012)).

FUNDING

Cohen's work was supported by National Institutes of Health grant 10T2TR003422-01 to Lawrence E. Hunter and by a fellowship from the National Library of Medicine to Kevin Bretonnel Cohen. Support for this work was provided by the National Center for Advancing Translational Sciences, National Institutes of Health, through the Biomedical Data Translator program, award number 10T2TR003422-01. Any opinions expressed in this document are those of the Translator community at large and do not necessarily reflect the views of NCATS, individual Translator team members, or affiliated organizations and institutions. The

[7] https://publicationethics.org/resources/discussion-documents/authorship, accessed 2023-01-24.

Table 5. Author contributions

	KBC	You	LEH
Conceived idea	*		
Wrote first draft	*		
Participated in analysis	*	*	*
Critically reviewed one or more drafts	*	*	*
Agrees to be accountable for the content of the work	*	*	*
Approved final version	*	*	*
Approved submission to this journal	*	*	*
Agrees with the conclusions	*	*	*
Agrees with the list and order of authors	*	*	*

TRANSLATOR publication committee reviewed the paper for compliance with funding acknowledgment and author inclusion guidelines.

ACKNOWLEDGMENTS

- Pengcheng Lu's hard-earned discovery that document separators are sometimes missing from our hard-earned CRAFT corpus coreference annotations in CONLL format, which trips up the CONLL scorer, guilted us into working even harder on the topic of this paper.
- Bill Baumgartner Jr., Mike Bada, Arrick Lanfranchi, Martha Palmer, Karin Verspoor, and certainly others whose names we have forgotten participated in unforgotten discussions of coreference resolution at various stages of our work on the topic.
- Atul Gawande's surprisingly fascinating book *The Checklist Manifesto* was surprisingly lively inspiration for the surprisingly novel topic of our paper.
- Dina Demner-Fushman, Pierre Zweigenbaum, Patrick Paroubek, Anna Rogers, Aurélie Névéol, and Karën Fort influenced our thinking about reproducibility and about publication practices quite a bit.
- Massimo Poesio encouraged the effort by responding with a smiley-face to the examples of bad coreference resolution reproducibility checklists.
- Zeus the Labrador Retriever comforted author KBC during a bout of post-COVID pertussis, much of which he spent thinking about this topic.
- We still think fondly of Janet Hitzeman.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found on GitHub at [https://github.com/KevinBretonnelCohe

REFERENCES

Aktaş, B., Solopova, V., Kohnert, A., and Stede, M. (2020). Adapting coreference resolution to Twitter conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2454–2460

Aloraini, A. and Poesio, M. (2021). Data augmentation methods for anaphoric zero pronouns. In *Proc. Computational Models of Reference, Anaphora and Coreference*

Alves, G. Â. d. S., Coêlho, J. F., and Leitão, M. M. (2021). Coreferential processing in elderly with and without Alzheimer's disease. In *CoDAS* (SciELO Brasil), vol. 33

Apostolova, E. and Demner-Fushman, D. (2009). Towards automatic image region annotation-image region textual coreference resolution. In *Proceedings of Human Language Technologies: The 2009*

Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. 41–44

Apostolova, E., Tomuro, N., Mongkolwat, P., and Demner-Fushman, D. (2012). Domain adaptation of coreference resolution for radiology reports. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. 118–121

Applebaum, A. (2018). Putin's grand strategy. *South Central Review* 35, 22–34

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 555–596

Bergsma, S., Lin, D., and Goebel, R. (2008). Distributional identification of non-referential pronouns. In *Proc. Association for Computational Linguistics*. 10–18

Bianchi, F. and Hovy, D. (2021). On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3895–3901

Bodnari, A., Szolovits, P., and Uzuner, Ö. (2012). MCORES: a system for noun phrase coreference resolution for clinical records. *Journal of the American Medical Informatics Association* 19, 906–912

[Dataset] Branco, A., Cohen, K. B., Vossen, P., Ide, N., and Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: Introducing an LRE special section

Cao, Y. T. and Daumé III, H. (2020). Toward gender-inclusive coreference resolution. In *Proc. Association for Computational Linguistics*

Caporaso, J. G., Deshpande, N., Fink, J. L., Bourne, P. E., Cohen, K. B., and Hunter, L. (2008). Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. In *Pacific Symposium on Biocomputing*. 640–651

Catoire, P. (2022). The humanitarian aspects of the Russian-Ukrainian war as seen through the eyes of a French volunteer. *European Journal of Emergency Medicine* 1, 158–159

Chen, B., Su, J., Pan, S. J., and Tan, C. L. (2011). A unified event coreference resolution by integrating multiple resolvers. In *Proc. International Joint Conference on Natural Language Processing*. 102–110

Chhugani, K., Frolova, A., Salyha, Y., Fiscutean, A., Zlenko, O., Reinsone, S., et al. (2022). Remote opportunities for scholars in Ukraine. *Science* 378, 1285–1286

Choi, M., Liu, H., Baumgartner, W., Zobel, J., and Verspoor, K. (2015). Integrating coreference resolution for BEL statement generation. In *Proceedings of the fifth BioCreative challenge evaluatio workshop. Sevilla, Spain*

Choi, M., Liu, H., Baumgartner, W., Zobel, J., and Verspoor, K. (2016a). Coreference resolution improves extraction of Biological Expression Language statements from texts. *Database* 2016

Choi, M., Verspoor, K., and Zobel, J. (2014). Evaluation of coreference resolution for biomedical text. In *MedIR@ SIGIR*

Choi, M., Zobel, J., and Verspoor, K. (2016b). A categorical analysis of coreference resolution errors in biomedical texts. *Journal of biomedical informatics* 60, 309–318

Chowdhury, M. F. M. and Zweigenbaum, P. (2013). A controlled greedy supervised approach for co-reference resolution on clinical text. *Journal of biomedical informatics* 46, 506–515

Cohen, K., Névéol, A., Xia, J., Hailu, N., Hunter, L., and Zweigenbaum, P. (2017a). Reproducibility in biomedical natural language processing. In *Proc. AMIA Annual Symposium*

Cohen, K. B. (forthcoming, 2023). *Writing about data science research: With examples from machine learning and natural language processing* (Cambridge University Press)

Cohen, K. B., Fort, K., Adda, G., Zhou, S., and Farri, D. (2016a). Ethical issues in corpus linguistics and annotation: Pay per hit does not affect effective hourly rate for linguistic resource development on amazon mechanical turk. In *Proc. Language Resources and Evaluation* (NIH Public Access), vol. 2016, 8

Cohen, K. B., Lanfranchi, A., Choi, M. J.-y., Bada, M., Baumgartner, W. A., Panteleyeva, N., et al. (2017b). Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics* 18, 1–14

Cohen, K. B., Rogers, A., et al. (2020). Reproducibility in biomedical natural language processing: A FAIR approach to what we need to know. *Proc. American Medical Informatics Association*

Cohen, K. B., Xia, J., Roeder, C., and Hunter, L. E. (2016b). Reproducibility in natural language processing: a case study of two R libraries for mining PubMed/MEDLINE. In *Proc. Language Resources and Evaluation* (NIH Public Access), vol. 2016, 6

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., et al. (2018). Three dimensions of reproducibility in natural language processing. In *Proc. Language Resources and Evaluation* (NIH Public Access), vol. 2018, 156

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*, vol. 139 (MIT press Cambridge, MA)

Crane, M. (2018). Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics* 6, 241–252

Cybulska, A. and Vossen, P. (2015). Translating granularity of event slots into features for event coreference resolution. In *Proc. EVENTS: Definition, Detection, Coreference, and Representation*. 1–10

Dunnigan, J. F. (2003). *How to make war: a comprehensive guide to modern warfare in the twenty-first century* (Harper Collins)

Dyer, C. (2016). Junior doctor is suspended for citing colleagues on falsified research without their knowledge. *British Medical Journal*

Fang, B., Baldwin, T., and Verspoor, K. (2022). What does it take to bake a cake? The RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*. 3481–3495

Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P., and Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. In *Proc. Association for Computational Linguistics*. 1691–1701

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics* , 413–420

Gîfu, D. and Cioca, L.-I. (2017). Detecting bridge anaphora. *International Journal of Computers, Communications, and Control* 12, 217–226

Gîfu, D. and Iliescu, A. (2015). Analysis of bridge anaphora across novel. *Procedia–Social and Behavioral Sciences* 180, 1474–1480

Gîfu, D. and Onofrei, M. (2017). Detecting bridge anaphora in novels. In *21st International Conference on Control Systems and Computer Science (CSCS)* (IEEE), 553–558

Gordin, M. D. (2012). *The pseudoscience wars: Immanuel Velikovsky and the birth of the modern fringe* (University of Chicago Press)

Grouin, C., Dinarelli, M., Rosset, S., Wisniewski, G., and Zweigenbaum, P. (2011). Coreference resolution in clinical reports-the limsi participation in the i2b2/va 2011 challenge. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*

Gu J, S. S., Sun R and Z, Y. (2015). Retraction. *OncoTargets and Therapy*

Hirschman, L. and 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*

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua* 44, 311–338

Jackson, P. and Moulinier, I. (2002). *Natural language processing for online applications* (John Benjamins Philadelphia)

Jauhar, S. K., Guerra, R., Pellicer, E. G., and Recasens, M. (2015). Resolving discourse-deictic pronouns: A two-stage approach to do it. In *Proc. Lexical and Computational Semantics*. 299–308

Johnson, H. L., Bretonnel Cohen, K., and Hunter, L. (2007). A fault model for ontology mapping, alignment, and linking systems. In *Pacific Symposium on Biocomputing*. 233–244

Joshi, M., Levy, O., Zettlemoyer, L., and Weld, D. (2019). BERT for coreference resolution: Baselines and analysis. In *Proc. Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*

Kantor, B. and Globerson, A. (2019). Coreference resolution with entity equalization. In *Proc. Association for Computational Linguistics*

Kilicoglu, H. and Demner-Fushman, D. (2014). Coreference resolution for structured drug product labels. In *Proceedings of BioNLP 2014*. 45–53

Kilicoglu, H. and Demner-Fushman, D. (2016). Bio-SCoRes: A smorgasbord architecture for coreference resolution in biomedical text. *PLOS ONE* 11, e0148538

Kim, J.-D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., and Yonezawa, A. (2012). The GENIA event and protein coreference tasks of the BioNLP shared task 2011. In *BMC bioinformatics* (BioMed Central), vol. 13, 1–12

Krippendorff, K. (1989). Content analysis

Kübler, S. and Zhekova, D. (2011). Singletons and coreference resolution evaluation. In *Proc. Recent Advances in Natural Language Processing*

Lang, J., Qin, B., Liu, T., and Li, S. (2007). (a review of textual coreference resolution research). *Journal of Chinese Language and Computing* 17, 227–253

Lapshinova-Koltunski, E., Ferreira, P. A., Lartaud, E., and Hardmeier, C. (2022). ParCorFull2.0: A parallel corpus annotated with full coreference. In *Proc. Language Resources and Evaluation*. 805–813

Lapshinova-Koltunski, E., Krielke, M.-P., and Hardmeier, C. (2020). Coreference strategies in English-German translation. In *Proc. Computational Models of Reference, Anaphora and Coreference*. 139–153

Lavergne, T., Grouin, C., and Zweigenbaum, P. (2015). The contribution of co-reference resolution to supervised relation detection between bacteria and biotopes entities. *BMC bioinformatics* 16, 1–17

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proc. Empirical Methods in Natural Language Processing*

Markert, K. and Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics* 31, 367–402

McCawley, J. D. (1976). Notes on Jackendoff's theory of anaphora. *Linguistic Inquiry* 7, 319–341

[Dataset] McCawley, J. D. (1983). Lingua mentalis: The semantics of natural language

McCawley, J. D. (1998). *The syntactic phenomena of English* (University of Chicago Press)

[Dataset] McShane, M. and Nirenburg, S. (2021). Basic coreference resolution

Mieskes, M. (2017). A quantitative study of data in the NLP community. In *Proceedings of the first ACL workshop on ethics in natural language processing*. 23–29

Mieskes, M., Fort, K., Névéol, A., Grouin, C., and Cohen, K. B. (2019a). Community perspective on replicability in natural language processing. In *Proc. Recent Advances in Natural Language Processing*. 768–775

Mieskes, M., Fort, K., Névéol, A., Grouin, C., and Cohen, K. B. (2019b). NLP community perspectives on replicability. In *Recent Advances in Natural Language Processing*

Mitkov, R. (2014). *Anaphora resolution* (Routledge)

Nguyen, N., Kim, J.-D., Miwa, M., Matsuzaki, T., and Tsujii, J. (2012). Improving protein coreference resolution by simple semantic classification. *BMC Bioinformatics* 13, 1–12

Okasha, S. (2002). *Philosophy of science: A very short introduction*, vol. 67 (Oxford Paperbacks)

Olex, A. L. and McInnes, B. T. (2021). Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be. *Journal of Biomedical Informatics* 118, 103784

Pedersen, T. (2008). Empiricism is not a matter of faith. *Computational Linguistics* 34, 465–470

Peled, Y. (1990). Non-referential pronouns in topic position in Medieval Arabic grammatical theory and in modern usage. *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 140, 3–27

Plokhy, S. (2015). *The gates of Europe: A history of Ukraine* (Basic Books)

Plokhy, S. (2018). The return of the empire: The Ukraine crisis in the historical perspective. *South Central Review* 35, 111–126

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to resolve bridging references. In *Proc. Association for Computational Linguistics (ACL-04)*. 143–150

Poibeau, T. (2003). *Extraction automatique d'information: Du texte brut au web sémantique* (Hermes Science)

Poibeau, T. (2011). *Traitement automatique du contenu textuel* (Lavoisier)

Popper, K. (1953). Science: Conjectures and refutations. *Conjectures and Refutations*

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proc. Association for Computational Linguistics*

Ray, K. S., Zurn, P., Dworkin, J. D., Bassett, D. S., and Resnik, D. B. (2022). Citation bias, diversity, and ethics. *Accountability in Research* , 1–15

Recasens, M., de Marneffe, M.-C., and Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy* , 47–88

Rink, B., Roberts, K., and Harabagiu, S. M. (2012). A supervised framework for resolving coreference in clinical records. *Journal of the American Medical Informatics Association* 19, 875–882

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

Sedoc, J., Rogers, A., Rumshisky, A., and Tafreshi, S. (2021). Proc. Insights from negative results in NLP. In *Proc. Insights from Negative Results in NLP*

Steinberger, F. (2022). The normative status of logic

Sukthanker, R., Poria, S., Cambria, E., and Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion* 59, 139–162

Tavares, G., Fajardo, I., Ávila, V., Salmerón, L., and Ferrer, A. (2015). Who do you refer to? How young students with mild intellectual disability confront anaphoric ambiguities in texts and sentences. *Research in developmental disabilities* 38, 108–124

Uppunda, A., Cochran, S., Foster, J., Arseniev-Koehler, A., Mays, V., and Chang, K.-W. (2021). Adapting coreference resolution for processing violent death narratives. In *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online: Association for Computational Linguistics), 4553–4559. doi:10.18653/v1/2021.naacl-main.361

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., and South, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association* 19, 786–791

Vos, R. A., Katayama, T., Mishima, H., Kawano, S., Kawashima, S., Kim, J.-D., et al. (2020). BioHackathon 2015: Semantics of data for life sciences and reproducible research. *F1000Research* 9

Wager, E. (2012). The Committee on Publication Ethics (COPE): objectives and achievements 1997–2012. *La Presse Medicale* 41, 861–866

Wang, C., Chen, S., and Wang, Z. (2022). Retraction note to: Electrophysiological follow-up of patients with chronic peripheral neuropathy induced by occupational intoxication with n-hexane. *Cell Biochemistry and Biophysics* 80, 267–267

Ware, H., Mullett, C. J., Jagannathan, V., and El-Rawas, O. (2012). Machine learning-based coreference resolution of concepts in clinical documents. *Journal of the American Medical Informatics Association* 19, 883–887

Wilkens, R., Oberle, B., Landragin, F., and Todirascu, A. (2020). French coreference for spoken and written language. In *Proc. Language Resources and Evaluation* (Marseille, France: European Language Resources Association), 80–89

Wongkoblap, A., Vadillo, M. A., Curcin, V., et al. (2021). Deep learning with anaphora resolution for the detection of Tweeters with depression: Algorithm development and validation study. *JMIR Mental Health* 8, e19824

Yin, K., DeHaan, K., and Alikhani, M. (2021). Signed coreference resolution. In *Proc. Empirical Methods in Natural Language Processing*. 4950–4961

Yu, J., Moosavi, N. S., Paun, S., and Poesio, M. (2021). Stay together: A system for single and split-antecedent anaphora resolution. In *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online: Association for Computational Linguistics), 4174–4184. doi:10.18653/v1/2021.naacl-main.329

Zhang, J. K., Botterbush, K. S., Bagdady, K., Lei, C. H., Mercier, P., and Mattei, T. A. (2022). Blast-related traumatic brain injuries secondary to thermobaric explosives: implications for the war in Ukraine. *World neurosurgery*

Zhekova, D. and Kübler, S. (2013). Machine learning for mention head detection in multilingual coreference resolution. In *Proc. Recent Advances in Natural Language Processing*. 747–754

Zheng, J., Chapman, W. W., Crowley, R. S., and Savova, G. K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics* 44, 1113–1122

Zimmermann, M. (2014). *Expletive and referential subject pronouns in Medieval French* (de Gruyter)

and (2015). (a survey of coreference resolution methods). 29, 1–12

, , , , , et al. (2019). (a survey of coreference resolution techniques). *()* 5, 16–35

5 APPENDIX A: CHECKLIST AND MINIMUM INFORMATION STANDARD FOR COREFERENCE RESOLUTION RESEARCH AND DEVELOPMENT, VERSION 1.0

For updates, see the repository at https://github.com/KevinBretonnelCohen/transcoref.

1. Data
   a. Document count
      - In training data INT
      - In devtest data INT
      - In blind test data INT
   b. Chain count
      - IDENT chain count INT
      - APPOS chain count INT
   c. Markables
      - Markable count INT
      - Markable distribution MODEL
   d. Structural
      - Sentence count INT
      - Token count INT
      - Word count INT
   e. Semantic
      - Semantic class count INT
      - Named entity count INT
2. Algorithm
   - Preprocessing
     a. Tokenizer REPOSITORY *or* NAME/VERSION NUMBER
     b. Sentence splitter REPOSITORY *or* NAME/VERSION NUMBER
     c. Paragraph splitter REPOSITORY *or* NAME/VERSION NUMBER
     d. Named entity taggers REPOSITORY *or* NAME/VERSION NUMBER
     e. Stemmer REPOSITORY *or* NAME/VERSION NUMBER
     f. POS tagger REPOSITORY *or* NAME/VERSION NUMBER
     g. Parser REPOSITORY *or* NAME/VERSION NUMBER
   - Algorithm
     - Algorithm category: 1+ of Machine learning (supervised), Machine learning (unsupervised), Sieve
     - Rules
       a. Rule types: 1+ of Binary (threshholded), Binary (deterministic)
       b. Prose description
       c. Pseudocode Code for/Grammar of
     - Machine learning

        a. Algorithm name LIST
        b. Randomization seed REPOSITORY
        c. Features LIST
    - Knowledge sources NAME/VERSION NUMBER
    - Error analysis categories LIST

3. Evaluation
    - Baseline LIST
    - Figures of merit LIST
4. Parameters
    - Experimental LIST
    - Configurational LIST
5. Availability
    a. Location of code REPOSITORY
    b. Location of knowledge sources REPOSITORY
    c. Location of data REPOSITORY
    d. Location of intermediate outputs LIST
    e. Location of data used for tables and figures LIST

APPENDIX B: EXPLICATION OF COREFERENCE RESOLUTION REPRODUCIBILITY CHECKLIST ITEMS

PROOFED TO HERE Here we explain how to use the items of the checklist in Appendix A.

- **Algorithm type:** A broad categorization of the approach, assuming a typology including something similar to (1) rule-based, (2) machine learning, (3) sieve, and (4) hybrid. The knowledge-based versus knowledge-free distinction is covered in a separate item. Examples of papers in the various categories include (1) rule-based: Choi et al. Choi et al. (2014, 2016b), the Hobbs algorithm Hobbs (1978); (2) machine learning: Ware et al. Ware et al. (2012), Rink et al. Rink et al. (2012); (3) hybrid: Kilicoglu and Demner-Fushman Kilicoglu and Demner-Fushman (2016). The reader will note that a paper's title might not be an accurate indicator of its actual approach.
- **Rule types:** A broad categorization of each role, assuming a typology similar to (1) weighted versus deterministic; (2) lexicalized or not. . .
- **Rules:** The rules themselves, with a prose description of the *intent* of the rules, and the code for implementing each of them.
- **Markable count:** The raw frequency of potentially coreferential items in the dataset. For example, if the dataset's annotation schema defines all noun phrases and all temporal expressions as potentially coreferential, then the count of markables is the sum of the count of noun phrases plus the count of temporal expressions. For examples of markable counts, see Poesio et al. (2004); Chen et al. (2011); Cybulska and Vossen (2015); Jauhar et al. (2015); Aktaş et al. (2020); Lapshinova-Koltunski et al. (2020); Wilkens et al. (2020); Aloraini and Poesio (2021); Yu et al. (2021) (**as of 2022-12-22**).
- **Markable distribution:** Any frequency information that is more granular than the overall count of markables in the corpus. Examples would be distribution across documents, across kinds of coreference, across kinds of reference, across genders, or anything else that is relevant to understanding the distribution of markables in a dataset with respect to how to interpret findings or predict generalizability (or lack thereof). For examples of reporting markable distributions, see Poesio et al. (2004); Chen et al. (2011); Cybulska and Vossen (2015); Jauhar et al. (2015); Rudinger et al. (2018); Aktaş et al. (2020); Lapshinova-Koltunski et al. (2020); Wilkens et al. (2020); Aloraini and Poesio (2021); Yu et al. (2021) (**as of 2022-12-22**).
- **Parsers/splitters/tokenizers/normalizers/stemmers/lemmatizers/taggers:** For third-party tools, identify them fully, including version numbers and any models used with them. For homegrown tools, give the location of the source code. If none are used, say so explicitly, rather than asking the reader to guess.
- **Counts of tokens/types/words/vocabulary size:** Define each of those terms—they are notoriously ambiguous. *Cite:* MEDINFO paper, Greffenstette paper, tokenization evaluation paper.
- **IDENT chain count:** The raw frequency of IDENT chains.
- **IDENT chain distribution:** for example, across lengths, across documents, whether or not singletons are included as one-item IDENT chains. . .
- **Knowledge sources:** Any external or internal source of linguistic or encyclopedic knowledge. Choi et al. (2014) have demonstrated that differences in knowledge sources can affect coreference resolution system performance even more than algorithmic differences. They include any statistical model learned from text; encyclopedic sources such as Wikipedia or a thesaurus; lexical sources such as WordNet;

Google counts Markert and Nissim (2005); corpora Markert and Nissim (2005); tokenizers and sentence splitters; even a trigram language model uses external knowledge of character encodings.

- **Code location:** Locations of code that implements the algorithm, varies configuration or experimental parameters, implements a processing pipeline, carries out an analysis, generates figures or tables…
- **Data location:** Locations of data used to evaluate and/or train the system; locations of any data-based knowledge sources, such as dictionaries, word lists, corpora…When evaluation or training data cannot be made publicly available, as is common in the medical domain, authors should be even more thorough than usual in documenting the population, sampling technique, inclusion criteria, and exclusion criteria, as well as describing the distributional characteristics of the dataset overall (e.g. number of health records, number of patients, distribution of documents per record…
- **Experimental parameters:** These are parameters that can be varied to test hypotheses or optimize system performance. For example: kinds of features, number of features, feature selection methods; tokenization and word normalization approaches; number of folds in cross-validation, seeds for randomization; sparseness of a Document-Term Matrix, minimum word count in a Document-Term Matrix…Experimental parameters typically either are intentionally varied from one run of the system to another, or are *deliberately* held constant from one run of the system to another.
- **Configurational parameters:** These are characteristics of the computational infrastructure. For example: computing platform, operating system version, number of CPUs used, versions of programming languages, versions of libraryies…Configurational parameters typically do not vary from one run of the system to another.
- **Figures of merit:** Specify all figures of merit; if one is used as the primary figure of merit, specify it. For the F-measure, specify the value of beta. If inter-annotator agreement is used as a figure of merit, specify how expected agreement was determined Krippendorff (1989); Artstein and Poesio (2008).
- **Baseline:** Describe the baseline measure completely. If the baseline is a third-party system, including previously published results, describe it in sufficient detail for the reader to be able to identify all differences between the baseline system and your system.
- **Error analysis categories:** Define error analysis categories and describe the inclusion and exclusion categories for each. Give examples, and the distribution across categories. Choi et al. (2016b) gives a well-developed taxonomy of coreference resolution errors. If categories are not mutually exclusive, give the overlaps.
- **Location of intermediate outputs:** Identify a repository containing the outputs of all steps of the processing. For example: the output of tokenization; the output of named entity recognition; the output of dictionary-cleaning…
- **Source of data for tables and figures:** Code for generating tables and figures; the data that populates those tables and figures…

APPENDIX B: A CASE STUDY IN USING THE CHECKLIST

This case study describes the process of applying the checklist during the course of development of a coreference resolution system by authors KBC, WABJr[8], and LEH. This is our preferred application of the checklist: during the course of research and development. As such, it served us as both a method of documentation *and* as a tool for planning.

As a tool for planning the research, it also made us aware very early of some weaknesses in the conception of the project. For example, while our reliance on the CONLL-X format makes the work much more easily repeatable by other labs, it also potentially limits the generalizability of the findings.

Algorithm category: The algorithm utilizes a sieve approach. The elements of the sieve are rule-based. It can make use of machine learning in its various components, and well might have (unbeknownst to us) during the preprocessing of the data. So, although the *algorithm* is rule-based, a resulting *system* might be a rule-based/machine learning hybrid Jackson and Moulinier (2002).

Rule types: Rules in the sieve can be deterministic, in which case it is a true sieve. However, the algorithm can also be modified for use as a rule-based classifier, in which case the rules can be deterministic or weighted.

The rules make use of two kinds of information: semantic, and syntactic. The semantic rules make use of named entities, labeled according to a set of biomedical ontologies. They operate at two levels. *Leaf-node semantic match* is a binary rule with value *1* if the anaphor and a candidate antecedent are labelled with the exact same leaf node. The rule has a value of *0* otherwise[9]. *Semantic class match* is a binary rule with value *1* if the anaphor and a candidate antecedent are labelled with elements of the same ontology. The rule has a value of *0* otherwise[10].

An example: if *amyotrophic lateral sclerosis* and *Charcot disease* are both labelled MESH D000690, they would match the *leaf-node semantic match* rule and the *semantic class match* rule. *Amyotrophic lateral sclerosis* labelled as MESH D000690 and *Charcot-Marie-Tooth disease* labelled as MESH D002607 would trigger the *semantic class match* rule, since they are both labelled as references to the Medical Subject Headings (MESH) taxonomy. They would not trigger the *leaf-node semantic match* rule, since they have different identifiers.

In contrast to the semantic rules, the syntactic rules rely on sentence position. In Hobbsian fashion, they ask whether or not (a) anaphor and candidate are in the same sentence, (b) anaphor and candidate are in adjacent sentences, (c) anaphor, candidate, or both are in sentence-initial or sentence-final position.

A third type of rule is purely orthographic, so I'm not sure what it counts as. Two rules ask whether or not (a) an anaphor and potential candidate are exact string matches, and (b) an anaphor and candidate are string matches if case-toggled. For example, *Zeus* and *Zeus* (referring to a Drosophila male fertility gene, flybase.org/reports/FBgn0032089.html) are exact string matches. In contrast, *Zeus* and *zeus* are string matches if case-toggled, but not otherwise[11].

Baseline(s): We hate the use of previously published results as a baseline, although we did include them in discussions of related literature. In the case of the CRAFT data, the only previously published results are

[8] Bill doesn't actually want to coauthor—no time.

[9] In the case of weighted rules, the values are some very large number, or 0.

[10] In the case of weighted rules, the values are some mid-sized number, or 0.

[11] Is this part of *Rule types*, or does it belong in the list of rules?

the baseline system in [Bill's shared task paper]. Instead, we used what Resnick and Lin have called *xxx baselines:*

1. Single rule: analogous to the xxxx "single feature" baseline. This is a kind of ablation experiment (see Cohen (1995) and Cohen (forthcoming, 2023) for the rationale behind it). Based on previous work on the CRAFT corpus, we expected the exact string match rule to be a non-trivial baseline.
2. Because the individual rules include occasionally high-performing simple syntactic rules (e.g. *closest preceding noun phrase*), the single rule baseline itself includes a number of strong candidates for baselines.

Figures of merit: We used all of the figures of merit calculated by the CONLL-X scoring code, viz. xxx, xxx, and xxx. This could be criticized as throw-them-all-against-the-wall-and-see-where-we-score-highest, but in fact we used all of them to test robustness of the results across different metrics.

Error analysis categories:

1. Is the error due to a software bug?
2. Is the error due to a design bug?
3. Is the error due to preprocessing?
4. Is the error related to a semantic rule?
5. Is the error related to a syntactic rule?

Location of intermediate outputs: Repository https://github.com/KevinBretonnelCohen/transcoref, spreadsheets xxx, xxx, and xxx.

Source of data for tables and figures: Recorded in the R script xxxx.Rmd in repository https://github.com/KevinBretonnelCohen/transcoref

Observations on using the checklist

In the course of writing the description of the algorithm, we produced what later became a substantial portion of a separate paper's *Methods* section.

We originally conceived of the *configurational parameter* versus *experimental parameter* contrast as being between parameters that are fixed across all system runs (e.g. the processor in KBC's laptop) and parameters that are varied (e.g. whether rules are deterministic or weighted). In the course of filling out the checklist, we realized that outputs for which we measure variability, such as dispersion across figures of merit, should probably also be considered an experimental parameter, since it produces a *result* that yields a *finding* that supports or fails to support some *result*, which may or may not contribute to a specific *conclusion*. As the reader will note, this is a fine example of the need to think about reproducibility along a variety of dimensions or levels…