NARC – Norwegian Anaphora Resolution Corpus

Petter Mæhlum,[1] Dag Haug,[2] Tollef Jørgensen,[3] Andre Kåsen,[4] Anders Nøklestad,[5] Egil Rønningstad,[1] Per Erik Solberg,[4] Erik Velldal,[1] and Lilja Øvrelid[1]

[1]University of Oslo, Language Technology Group
[2]Department of Philosophy, Classics, History of Arts and Ideas, University of Oslo
[3]Department of Computer Science, Norwegian University of Science and Technology
[4]The Norwegian Language Bank, National Library of Norway
[5]Department of Linguistics and Scandinavian Studies, University of Oslo

Abstract

We present the Norwegian Anaphora Resolution Corpus (NARC), the first publicly available corpus annotated with anaphoric relations between noun phrases for Norwegian[1]. The paper describes the annotated data for 326 documents in Norwegian Bokmål, together with inter-annotator agreement and discussions of relevant statistics. We also present preliminary modelling results which are comparable to existing corpora for other languages, and discuss relevant problems in relation to both modelling and the annotations themselves.

1 Introduction

Coreference resolution (CR) is a central NLP task which enables a wide range of applications aiming to extract and aggregate various types of information from text, e.g. relations, events and opinions. While a number of datasets for CR have been developed for a range of different languages, no such openly available dataset is currently available for Norwegian.

In this paper, we describe the annotation of the Norwegian Anaphora Resolution Corpus (NARC). The annotation effort enriches the existing annotation of the Norwegian Dependency Treebank (NDT) (Solberg et al., 2014), which has been converted to the Universal Dependencies standard (Øvrelid and Hohle, 2016; Velldal et al., 2017) and has further been annotated with named entities in a separate effort, resulting in the NorNE dataset (Jørgensen et al., 2020). Norwegian has two written standards: Bokmål and Nynorsk, and the dataset consists of 300,000 tokens from each.[2]

The paper is organized as follows: we start out by reviewing related work, then we describe the annotation effort, summarize the annotation procedure, explain guidelines developed for the project and the inter-annotator agreement scores. Finally, corpus statistics and initial experiments with Norwegian CR are presented, before concluding the paper.

[1] https://github.com/ltgoslo/NARC

[2] We here focus on the annotation of the Bokmål part of the NDT, however, annotation of the Nynorsk part of the treebank follows the same guidelines and is currently close to completion. The final version of the corpus will include statistics and data for both written standards.

2 Related Work

In this section we review some related work, both in terms of existing datasets with coreference annotation and research on coreference modelling based on these datasets.

2.1 Datasets

Early datasets for CR were MUC (Grishman and Sundheim, 1996) and ACE (Doddington et al., 2004), which enabled considerable research on this task, further spurred by the CoNLL-2011 and 2012 shared tasks on CR (Pradhan et al., 2011, 2012) based on the widely used OntoNotes dataset (Weischedel et al., 2011).

There are now a wide range of annotated coreference datasets. A majority of these are in English, such as Quiz Bowl Coreference (Guha et al., 2015), Character Identification, (Chen and Choi, 2016), WikiCoref (Ghaddar and Langlais, 2016), GUM (Zeldes, 2017), BASHI (Rösiger, 2018), PreCo (Chen et al., 2018), GAP (Webster et al., 2018), ARRAU (Uryupina et al., 2020) and LitBank (Bamman et al., 2020).

There is also a growing number of non-English corpora being made available to the research community, including datasets for Catalan/Spanish (Recasens and Martí, 2010), Czech (Nedoluzhko et al., 2016), Danish (Korzen and Buch-Kromann, 2011), Dutch (Hendrickx et al., 2008), German (Lapshinova-Koltunski et al., 2018; Bourgonje and Stede, 2020), Hungarian (Vincze et al., 2018), Lithuanian (Žitkus and Butkiene, 2018), Polish (Ogrodniczuk et al., 2016) and Russian (Toldova and Ionov, 2017). The recent Universal Anaphora

initiative[3] constitutes an important step towards the harmonization of different annotation standards for corpora annotated with various types of anaphoric information. A dataset of particular relevance to the current work is BREDT (Borthen et al., 2007) – annotated with coreference and other anaphoric relations in Norwegian. BREDT covers in total 12 different kinds of relations, all linguistically motivated. The data has been previously used both to test a rule-based (Holen, 2007) and a machine learning-based system (Nøklestad, 2009) for Norwegian CR. Unfortunately, however, the BREDT dataset is not openly available.

2.2 Modelling

A variety of CR approaches have been published using the MUC and ACE datasets, ranging from linear programming, probabilistic and rule-based mention–pair models (Ng and Cardie, 2002; Luo et al., 2004; Culotta et al., 2007; Denis and Baldridge, 2007; Finkel and Manning, 2008; Haghighi and Klein, 2009). These datasets were of limited size, and Poon and Domingos (2008) found that unsupervised models were comparable to supervised models at the time – an important observation for low-resource languages. After the SemEval 2010 (Recasens et al., 2010) and CoNLL shared tasks (Pradhan et al., 2011, 2012), more extensive models were proposed, such as the ranking models by Björkelund and Farkas (2012); Durrett and Klein (2013), the sieve-based deterministic model by Lee et al. (2013) and other machine learning-based methods (Clark and Manning, 2015, 2016). Recent state-of-the-art models, however, such as those by Agarwal et al. (2019); Wu et al. (2020); Kantor and Globerson (2019); Xu and Choi (2020); Joshi et al. (2020); Kirstain et al. (2021) and Dobrovolskii (2021) have mostly been evaluated on the OntoNotes dataset. This is perhaps due to lack of compatability in terms of formats, annotation styles, and genres across datasets. Consequently, there are several concerns regarding real-world use of models that are not evaluated on other domains, especially regarding domain generalizability and robustness (Guha et al., 2015; Moosavi, 2020; Sukthanker et al., 2020). The same issues will likely translate to NARC, as data sources are limited (see Section 3.1). To tackle these issues, cross-domain adaptability will be a central topic for future evaluation.

[3] https://github.com/UniversalAnaphora

For computing preliminary benchmark results for NARC – as presented in Section 5 – we adopt the approach for word-level coreference resolution developed by Dobrovolskii (2021)[4]. Rather than directly predicting coreference links between word spans, the problem is split into two sub-tasks; first predicting coreference links between individual words, and then predicting the corresponding spans. This substantially reduces the computational complexity while still maintaining SotA performance when evaluated on OntoNotes for English, owing in particular to gains in recall (Dobrovolskii, 2021).

3 Annotation

We here detail the annotation effort and present the underlying data for annotation, the pre-annotation of markables, the annotated NARC mentions and relations, as well as the review and curation process and inter-annotator agreement.

3.1 Data source

As mentioned above, the underlying data for the annotation effort is the Norwegian Dependency Treebank (NDT), a richly annotated dataset (Solberg et al., 2014; Øvrelid and Hohle, 2016; Jørgensen et al., 2020). The original treebank contains manually annotated syntactic and morphological information for both varieties of written Norwegian – Bokmål and Nynorsk – comprising roughly 300,000 tokens of each and a total of around 600,000 tokens. The corpus contains a majority of news texts (comprising around 85% of the corpus), but also other types of texts, such as government reports, parliamentary transcripts and blog data.

3.2 Pre-annotation

In order to alleviate the annotators' job of locating potential mentions for coreference, we make use of the existing syntactic annotation of the data to perform a pre-annotation step. In particular, we formulate simple heuristics over parts-of-speech and dependency relations which derive noun phrases from the dependency syntax of the treebank. Using the dependency syntax, we extract all nominal heads that are either i) nouns (both common and proper nouns), ii) referential personal pronouns [5],

[4] Information on the modelling setup is available from the data repository.

[5] The NDT annotation identifies so-called formal subjects/objects, which are non-referential or expletive uses of the pronoun *det* 'it'.

iii) possessive pronouns, or iv) adjectives in a nominal syntactic function (subject, object or prepositional complement). The full NP is constructed by traversing all syntactic dependents of these nominal heads. For coordination, we extract the full coordinated phrase as well as potential markables for individual nominal conjuncts. The annotators are instructed to treat the pre-annotated markables as suggestions only, since the syntactic units do not always correspond to coreference mentions (Popel et al., 2021).

3.3 Annotation guidelines

The annotation guidelines were developed during an initial pilot phase, where the documents used for training of the annotators were annotated by two of the project PIs. The guidelines were based largely on the guidelines from Ontonotes and the previous Norwegian BREDT dataset, as described in section 2.1 above, and were continuously refined following discussions and inputs from the annotators. The full set of annotation guidelines are released along with the dataset.

3.4 NARC markables

The annotators are presented with the pre-annotated markables for annotation. As mentioned above, these include nouns, referential personal and possessive pronouns, as well as adjectives in a nominal function. Below we describe some of the specific cases regarding the annotation of markables in NARC.

Markable boundaries Compounding is highly productive in Norwegian and compounds are written as one word, e.g. *innebandylag* 'field hockey team'. Even so, markables in NARC always correspond to full tokens and are never sub-token units. Additional information that is often provided in parentheses behind a noun, e.g. *John (53)* is part of the noun phrase and therefore also part of the markable in NARC. Both relative clauses and appositions are also included in the span of the markable that they modify.

Nested markables NARC allows for nested markables, i.e., when a nominal markable is contained within a larger markable. When considering pre-annotated markables that were nested, the annotators were instructed to assess whether it is possible for the individual nominals making up the larger markable to have a reference that is independent of the markable as a whole. Only in cases where this is in fact possible should nesting of markables be allowed. Proper nouns are always considered to be atomic and they are not annotated as nested even if it is possible to identify composite proper nouns within the names, such as e.g., *Oslo* in the proper name *University of Oslo*. This treatment is also in line with the flat annotation of such names in both the original treebank (Solberg et al., 2014) and NorNE (Jørgensen et al., 2020), the named entity annotated version of the treebank as described above.

3.5 NARC relations

Three relations are used in NARC: COREFERENCE, BRIDGING and SPLIT-ANTECEDENT.[6] In the following we describe the annotation of these relations in NARC, relating to annotation efforts for other datasets wherever possible.

3.5.1 Coreference

COREFERENCE is the relation reserved for coreferring markables. The annotation guidelines are to a large extent based on those of OntoNotes (Weischedel et al., 2011). Two broad categories of coreferring expressions are recognised in NARC: *anaphors* and what we might called *repeated coreferring entities*. Anaphors, or anaphoric expressions, usually need to be resolved to an antecedent to be interpreted. This includes third person pronouns and possessive determiners such as *hun*, 'she' and *hans*, 'his', but also definite nouns such as *bilen*, 'the car'. The second category, *repeated coreferring entities*, are markables such as proper names and first and second personal pronouns, which are not inherently anaphoric, but which still can corefer with a markable in the previous text. Indefinite nominals, including many quantified expressions, are not assumed to be coreferent with a markable in the previous text, but they can be antecedents of anaphors.

Markables are generally linked to the nearest coreferring markable to the left. Figure 1 illustrates this: The spans marked with boxes are the markables of the text. The pronoun *han*, 'he', has a coreferent relation to *Henrik Bjørnstad* in the preceding sentence. This is, however, not always the case. In some instances, pronouns may resolve to markables that *follow* rather than precede – a

[6] Unlike OntoNotes, there is no relation for appositives (BBN Technologies, 2007, 1.2). Instead, the adjacent, coreferring nouns in an appositive construction are taken as part of the same markable span.

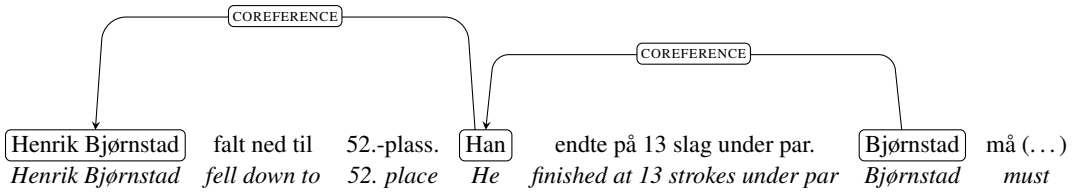

Figure 1: Example of a coreference relation in NARC.

phenomenon called *cataphora*. In cataphoric expressions, the markable is linked to the nearest antecedent to the right. This is shown in Figure 2, where the antecedent of the pronoun appears after the reference.

OntoNotes permits certain markables which are neither nominals nor determiners. Firstly, coreference relations are drawn between anaphoric expressions and verbs in OntoNotes. This means that e.g. event-denoting definite descriptions such as *the large growth* can refer back to a verb such as *grew*, which thereby becomes a markable. In NARC, however, all markables are nominal. Secondly, temporal adverbials such as *now* and *then* may participate in coreference chains in OntoNotes (BBN Technologies, 2007, 1.1.4; 2.8), whereas we only annotate temporal expressions that are nominal.

In NARC, we have chosen not to include verbs and adverbs in the set of possible markables. While this may leave certain anaphoric markables without an antecedent, it makes the annotation task easier and removes a potential source of inconsistencies. It is, for example, not always clear if the actual antecedent of an anaphoric expression is a verb or an entire proposition.

3.5.2 Split antecedent

The anaphoric possibilities of plural pronouns and definite nouns are a bit broader than for singular anaphors. They may corefer with a plural nominal or a coordinated structure in the textual context, in which case it is annotated with a COREFERENCE relation in NARC. Quite often, however, the reference of the plural anaphor is not coreferent with one single markable, but rather has multiple 'partial' antecedents in the discourse. Such cases are treated differently in different datasets. In OntoNotes, they are not annotated at all. In the ARRAU corpus, they are handled as a kind of bridging (Uryupina et al., 2020, pp. 106-107). In NARC, we use a special relation in such cases: SPLIT-ANTECEDENT. A split antecedent relation is drawn from the anaphor to each of its partial antecedents. This is shown in figure 3.

3.5.3 Bridging

BRIDGING indicates an anaphoric relation between two markables that are *not* coreferent, but associated in such a way that the correct identification of the anaphoric referent requires that the hearer establishes the relation to the antecedent. For example, in Figure 4, *rattet* 'the wheel' refers to the steering wheel of the mentioned car, *den gullfargede Roveren* 'the gold-colored Rover'. Typical relations involved in bridging are part–whole relations and various types of possession (Clark, 1977). Bridging can also involve verbal antecedents, where a following definite nominal is understood to have filled a particular thematic role: *John was murdered yesterday. The knife laid nearby*. In line with our decision to exclude verbal antecedents, we do not annotate these.

There are fewer corpora with bridging annotation compared to those which annotate coreference. For example, OntoNotes does not include bridging annotation, although two later efforts, ISnotes (Markert et al., 2012) and BASHI (Rösiger, 2018), each added this for 50 WSJ articles from OntoNotes. The ARRAU corpus (Uryupina et al., 2020) also includes bridging annotations.

Bridging is a complex phenomenon, with several sub-types and no established annotation standard; see the discussion in Roesiger et al. (2018). For our purposes, we adopted a very simple heuristic: when encountering a definite NP, annotators were asked first to look for a coreferent antecedent. If there is none, they should look for a related but not coreferent NP (e.g. bearing a part–whole relation or a possessive relation) and consider whether that related NP explains the use of the definite article by imagining the text without the antecedent. If this makes the definite infelicitous, it should be marked with BRIDGING. We make no attempt to identify sub-types of bridging.

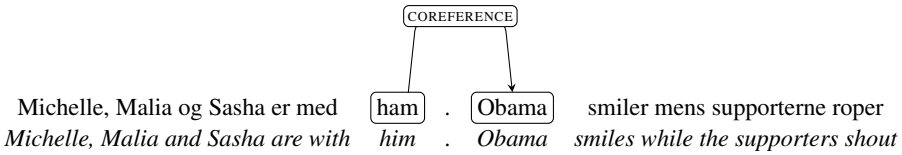

Figure 2: Example of a cataphora relation in NARC.

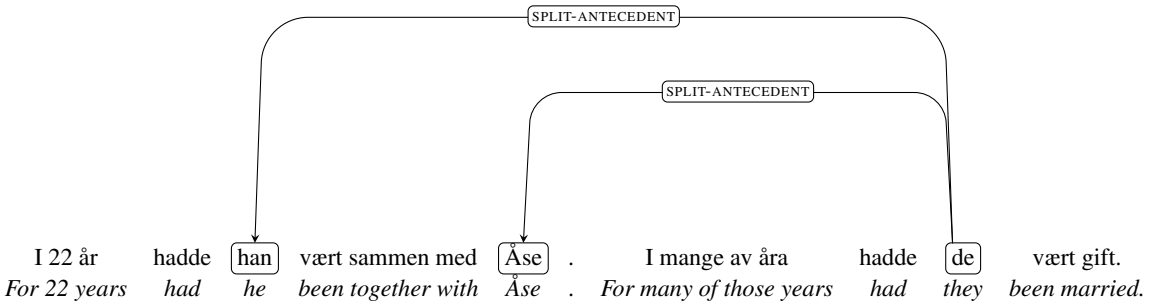

Figure 3: Example of a split antecedent relation in NARC.

3.6 Annotation Procedure

Annotation was performed using the Brat annotation tool (Stenetorp et al., 2012). Six students with a background in NLP and linguistics annotated the Norwegian Bokmål part of the corpus. The students received financial remuneration for their annotation work. All annotators completed an initial training round where they were tasked with annotation of the same set of documents, followed by a round of discussion and consolidation, along with updates to the annotation guidelines.

Due to restrictions in the annotation software, and the time needed to annotate, documents of over 150 sentences in length were split into smaller sections and annotated separately. All other documents were sorted into groups of 10 which were balanced according to length to ensure a constant workload for the annotators across the annotation period. During weekly meetings, the annotators had the opportunity to discuss challenges encountered when annotating or unclarities in the guidelines, so that these could be resolved, and the guidelines updated. Note that the documents set aside for measuring inter-annotator agreement were exempt from these discussions.

3.6.1 Review and Curation

Following the initial annotation process, all documents were re-annotated in one of two ways. The documents annotated by a single annotator were *reviewed* by a second annotator in a subsequent step. In this case, the second annotator only corrected errors from the first annotation round. In the case of documents annotated for inter-annotator agreement, a third annotator would base the curation of the document on one annotation, and then make changes based on the other, ensuring that both annotations are taken into account, while at the same time making sure there are no errors. Although both addition and removal of annotations were seen in the review process, the average changes were positive in all cases. These differences in numbers for the relations are summarized in Table 1.

Markables	Relations	Coref.	Bridg.
1.7	5.9	4.5	1.5

Table 1: Differences in numbers before and after review. Numbers are average differences between documents.

3.7 Inter-Annotator Agreement

59 documents, divided into 5 groups of 10 and one group of 9, were set aside for inter-annotator agreement towards the end of the annotation period. Each document group was annotated by two annotators. All annotators annotated at least one group for IAA, while some annotated more, due to differences in capacity. These documents were chosen as they are believed to represent a point in time where annotators should be familiar with the guidelines and the annotation task. In order to get a reliable indication of which areas are the most problematic, we look at agreement scores for different components separately. We follow Nedoluzhko et al. (2016) in using an F_1 score to look at the

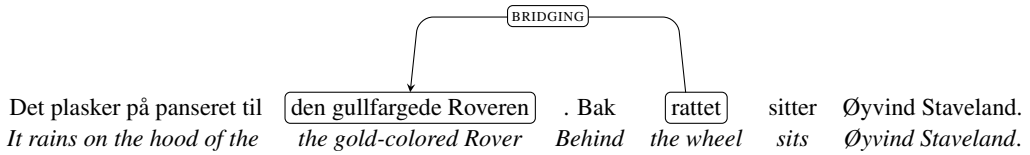

Figure 4: Example of a bridging relation in NARC.

agreement for all relations, and Cohen's κ for the specific labels. We also use the F_1 score for the markable agreement, following Kopeć and Ogrodniczuk (2014).

We see that annotators largely agree on the markables in the document, with some minor differences. On average, there was a difference of 2.2 markables per document, giving an F_1 score of 0.99. We see this as a confirmation that the pre-annotation provided a satisfactory basis for the annotation. Notably, 17% of the disagreement is due to the word *seg* 'oneself', which was known to fall outside of the pre-annotations in certain cases.

For the relations, we measure an overall F_1 score of 0.83. We see that although annotators tend to agree on many relations, there was still disagreement that had to be addressed during the review phase. When calculating the observed Cohen's κ, we follow Kopeć and Ogrodniczuk (2014), who notes that Cohen's κ must be calculated separately for each document, then averaged across documents in order to avoid including the probability of annotating across documents. The agreement score was calculated based on the markables that were already used in some relation by the annotators, and the relations for each annotator. The resulting values are presented in Table 2. We note that IAA scores are relatively high, especially for COREFERENCE. Both SPLIT-ANTECEDENT and especially BRIDGING have lower scores than COREFERENCE, but they also have fewer annotated examples. The low score for BRIDGING is also not surprising, based on the observation that this is a much more difficult annotation task.

4 Corpus statistics

Table 3 summarizes the most important statistics for the dataset [7]. We see that unsurprisingly the most common type of relation by far is the COREFERENCE relation, followed by BRIDGING and SPLIT-ANTECEDENT. However, some of these low numbers for BRIDGING can be explained by the difficulty of identifying bridging in the first place. We see that despite a large number of possible markables from the pre-annotation process, only 37% are used in relations. Relations are overwhelmingly anaphoric, with only 1.3% being cataphoric. As we do not pose any restrictions on how far back a relation can be drawn in a document, there are some relations with long edges. Looking at the distance based on tokens, the mode distance is 6, but the distribution has a long tail to the right with many long-distance relations. An example of this is in one of the documents with more than 150 sentences, where a relation was drawn from near the end of the document to an antecedent near the start, separating the elements of the relation by 5629 tokens. These cases do require that no relevant antecedent be mentioned in between. Due to the long tail, the average distance is 70.4, while the median is 19.0 for COREFERENCE. For BRIDGING, the average is 32.1 while the median is 16.0. Note that the annotators were told to think about whether the removal of the antecedent in a BRIDGING relation would change the viability of the definite form believed to have a BRIDGING relation. This might have caused an implicit restriction on bridging-relation lengths. For COREFERENCE there were no such restrictions, and annotators were asked to mark all COREFERENCE relations where possible. The median for the split antecedent relations is 22.0.

Clusters are collections of relations that have markables in common. The average length for COREFERENCE clusters is 4.7 tokens, while for BRIDGING it is 2 tokens. Most clusters, regardless of type, are of length 2, i.e. from a single antecedent to an anaphoric or bridging expression. Despite the low average, there are still some very long clusters.

Finally, we also analyzed the data to investigate what types of expressions occur as anaphoric expressions. As noted earlier, there are primarily two types of relations that fall under COREFERENCE.

[7] These statistics correspond to the first version of the NARC corpus. Subsequent releases of the dataset will contain the full Bokmål part of NDT as well as the Nynorsk part of the corpus.

	Overall F_1	Anaphor κ	Cataphor κ	Coref. κ	Bridging κ	Split Ant. κ
Scores	0.83	0.82	0.80	0.84	0.44	0.66

Table 2: IAA scores for the 59 documents annotated for agreement. The overall score is in F_1, while the others are represented by Cohen's κ, showing scores for specific directions (anaphor and cataphor) and labels (coreference, bridging, split antecedent).

Type	Value
Documents	326
Sentences	15125
Tokens	231363
Total markables	6979
Used markables	26005
Singletons	43788
Single word markables	34
Discontinuous markables	499
COREFERENCE relations	19420
BRIDGING relations	990
SPLIT-ANTECEDENT relations	292
COREFERENCE clusters	5350
BRIDGING clusters	962
Anaphor relations	20425
Cataphor relations	277
Sentences per document	46.4
Tokens per document	709.7
Markables per document	214
Avg. COREFERENCE cluster length	4.7
Avg. BRIDGING cluster length	2.0
Avg. COREFERENCE distance	70.4
Avg. BRIDGING distance	32.1
Avg. SPLIT-ANTECEDENT distance	53.9

Table 3: Counts and average values for some key statistics in the dataset. Singletons are markables that are not used in any relation. The last three values are the average distance between the antecedent and the referring expression in tokens.

The most common COREFERENCE expressions are pronouns, but both true anaphoric pronouns and pronouns referring to repeated entities are common. About 38% of all COREFERENCE relations are from a pronoun. As only third person pronouns and definite nouns can give rise to BRIDGING relations, this is naturally reflected in the types of expressions found. The most common is the pronoun *de* 'they', but another notable feature of the BRIDGING relations is that we see word forms such as *hodet* 'the head' *øynene* 'the eyes' *hånden* 'the hand' and *skuldrene* 'the shoulders' among the most common words. These are all typical of inalienable body parts, a type of bridging mentioned specifically in BREDT (Borthen et al., 2007).

5 Experiments

This section presents preliminary benchmarking experiments on the new dataset. Below we describe the distribution format of the data, the framework used for modelling and evaluation, and the results.

5.1 Format

Prior to modelling, the resulting files from the annotation tool (Brat) were converted to the format JSON Lines. This format has been common in coreference modelling since Lee et al. (2018) described the minimization process from the OntoNotes' CoNLL-format to JSON Lines, stripped of PoS-tags, lemmas and word sense information. For NARC, the annotations represent tokens structured in sentences along with coreference mention clusters, similar to LitBank (Bamman et al., 2020), GUM (Zeldes, 2017) and PreCo (Chen et al., 2018). Singleton mentions, i.e. markables not included in a coreference chain (see Table 3), have been discarded from the post-processing tasks, but may be used separately to model the impact of a separate mention detection system, as briefly studied by Chen et al. (2018), or a variation of the mention-ranking systems by Clark and Manning (2016). The dataset will include the data as JSON Lines and CoNLL, with and without singleton mentions. Furthermore, aligning NARC with the Norwegian Dependency Treebank (NDT), we will release the dataset in the CorefUD (Coreference Universal Dependencies) format, as described by Nedoluzhko et al. (2022).

5.2 Modelling framework

We apply the framework for word-level coreference resolution (wl-coref) developed by Dobrovolskii (2021), as mentioned in Section 2. This two-stage approach first predicts candidate antecedents for each token, before reconstructing the full spans by predicting the most likely start- and end-tokens in

Model	MUC			B^3			$CEAF_e$			LEA			CoNLL
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	Mean F_1
NorBERT2	**90.40**	79.35	84.52	**63.15**	**62.71**	**62.93**	**55.52**	33.54	41.82	**61.94**	**61.50**	**61.72**	**63.09**
XLM-R	84.97	**84.51**	**84.74**	61.09	49.09	54.44	51.17	**51.17**	**51.17**	58.87	47.11	52.34	**63.45**

Table 4: Evaluation of predictions on the held-out test split of NARC.

the same sentence. To create the required training data, the syntactic head for each annotated span is added to the dataset through the Norwegian parser available with spaCy[8]. On the basis of this, two training sets are created; one for predicting the word-level coreference links and one for predicting the corresponding spans (Dobrovolskii, 2021).

The original wl-coref system was trained with a 48 GB GPU resource. Our model was trained using a 40 GB GPU resource, which was sufficient to run the *base* model of XLM-RoBERTa with the same hyperparameters as Dobrovolskii (2021), but not the *large* version.

For evaluation we use the standard coreference metrics as computed by the CoNLL 2012 scoring script, including the MUC metric proposed by Vilain et al. (1995), B^3 as proposed by Bagga and Baldwin (1998), $CEAF_e$ as proposed by Luo (2005), and finally the aggregated score of Mean F_1 as proposed by Pradhan et al. (2012), referred to as the CoNLL-F_1. We also evaluate with the Link-based Entity-Aware metric (LEA) by Moosavi and Strube (2016), using standard settings for entity importance scores.

Training a model for Norwegian text limits the options for pretrained language models. We chose four transformer-based language models for introductory testing on a subset of the data, two Norwegian and two multilingual, namely NorBERT2 (Kutuzov et al., 2021), NB-BERT (Kummervold et al., 2021), XLM-RoBERTa (XLM-R base) (Conneau et al., 2020) and multilingual BERT (mBERT) (Devlin et al., 2019).

Dobrovolskii (2021) report the choice of pretrained language models to be important for the system's performance. They use large, monolingual versions of RoBERTa, SpanBERT and Longformer. Such models are presently not available for Norwegian.

5.3 Results

The four language models were evaluated using wl-coref. Based on these initial results, as seen in Table 5, the NorBERT2 and XLM-RoBERTa models were selected for further experimentation. We proceeded with fine-tuning the two models on the training set, comprising 80% of the data. Two other splits – dev and test – were used for evaluation and a held-out test set respectively. Results on the test set are shown in Table 4. The high MUC scores indicate that the model was able to properly group mention clusters. The somewhat lower recall scores shows that there are still some lacking clusters, regardless of the groups they were linked to. B^3 and $CEAF_e$ scores are significantly lower, meaning that while a lot of mentions were found, the models discovered fewer entities and was unable to correctly assign mention clusters. The LEA score also represents the lack of entity assignment within the discovered clusters, and the higher score compared to $CEAF_e$ of the NorBERT2 model is likely due to LEA supporting a weighted one-to-many assignment of clusters.

mBERT	**nb-BERT**	**nor-BERT2**	**XLM-R**
51.3	51.8	54.0	56.1

Table 5: The four preliminary selected pretrained language models and their F_1 scores according to the wl-coref evaluation.

Regardless, we find that the scores are comparable to existing work on CR, with the main difference being the MUC values scoring higher than current state-of-the-art models on the OntoNotes dataset. The reason for lower scores on the following metrics are, as discussed, likely due to issues with entity resolution and assignment, and this is thus an important takeaway for future work.

6 Conclusion

This paper has introduced a new corpus for coreference resolution: the Norwegian Anaphora Resolution Corpus (NARC). It is the first openly available corpus of this kind for Norwegian and represents the result of a large annotation effort which en-

[8] https://spacy.io/

riches the Norwegian Dependency Treebank (Solberg et al., 2014; Øvrelid and Hohle, 2016) with annotation of document-level coreference resolution, including the annotation of split antecedents and bridging. The paper has detailed the annotation effort, including a summary of guidelines, annotion procedure, inter-annotator agreement and resulting dataset statistics, as well as provided results from initial modelling experiments. While this paper focuses on the annotation of the Bokmål section of the corpus, the final corpus will contain the full treebank dataset, including also its Nynorsk sections, corresponding to the second written standard of Norwegian. NARC, including the annotation guidelines, will be made freely available[9]. It will further be aligned with the underlying treebank, allowing for smooth interaction with the other annotation layers such as PoS, dependency syntax and named entities, thus constituting a richly annotated resource for Norwegian NLP in the future.

Acknowledgements

We want to express our gratitude to the many annotators involved with annotating the datasets: Fredrik Aas Andreassen, Marie Emerentze Fleisje, Jennifer Juveth, Annika Willoch Olstad, Anne Oortwijn, Stian Ramstad, Lilja Charlotte Storset, Veronica Dahlby Tveitan and Alexandra Wittemann. We are grateful for the initial funding from Teksthub, and to Språkbanken for the main funding of the project.

References

Oshin Agarwal, Sanjay Subramanian, Ani Nenkova, and Dan Roth. 2019. Evaluation of named entity coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–7, Minneapolis, USA. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation, Workshop on Linguistics Coreference*, pages 563–566, Granada, Spain.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

[9] https://github.com/ltgoslo/NARC

BBN Technologies. 2007. Co-reference guidelines for English OntoNotes version 7.0. Technical report, BBN Technologies.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 49–55, Jeju Island, Korea. Association for Computational Linguistics.

Kaja Borthen, Lars G. Johnsen, and Christer Johansson. 2007. Coding anaphor-antecedent relations; the annotation manual for bredt. In *Proceedings from the first Bergen Workshop on Anaphora Resolution (WAR1)*, pages 86–111. Cambridge Scholars Publishing.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France. European Language Resources Association.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles. Association for Computational Linguistics.

Herbert H Clark. 1977. Bridging. In Philip N. Johnson-Laird and Peter C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 311–326. Cambridge University Press, Cambridge.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 81–88, Rochester, New York. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243, Rochester, New York. Association for Computational Linguistics.

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

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of ACL-08: HLT, Short Papers*, pages 45–48, Columbus, Ohio, USA. Association for Computational Linguistics.

Abbas Ghaddar and Philippe Langlais. 2016. Coreference in Wikipedia: Main concept resolution. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 229–238, Berlin, Germany. Association for Computational Linguistics.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, pages 466–471.

Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. 2015. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1108–1118, Denver, Colorado. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1152–1161, Singapore. ACL and AFNLP.

Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A coreference corpus and resolution system for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Gordana Ilić Holen. 2007. Automatic anaphora resolution for norwegian (ARN). In *Anaphora: Analysis, Algorithms and Applications*, pages 151–166, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating named entities for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 14–19, Online. Association for Computational Linguistics.

Mateusz Kopeć and Maciej Ogrodniczuk. 2014. *Inter-annotator Agreement in Coreference Annotation of Polish*, pages 149–158. Springer International Publishing, Cham.

Iorn Korzen and Matthias Buch-Kromann. 2011. Anaphoric relations in the copenhagen dependency treebanks. In *Proceedings of DGfS Workshop, Göttingen, Germany*, pages 83–98.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 135–142, Barcelona, Spain.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Nafise Sadat Moosavi. 2020. *Robustness in Coreference Resolution*. Ph.D. thesis, Neuphilologische Fakultät > Institut für Computerlinguistik – Heidelberg University, Heidelberg.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 169–176, Portorož, Slovenia. European Language Resources Association (ELRA).

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Peter Bourgonje, Silvie Cinková, Jan Hajič, Christian Hardmeier, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, M. Antònia Martí, Marie Mikulová, Maciej Ogrodniczuk, Marta Recasens, Manfred Stede, Milan Straka, Svetlana Toldova, Veronika Vincze, and Voldemaras Žitkus. 2022. Coreference in universal dependencies 1.0 (CorefUD 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Anders Nøklestad. 2009. *A machine learning approach to anaphora resolution including named entity recognition, PP attachment disambiguation, and animacy detection*. Ph.D. thesis, University of Oslo.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2016. Polish coreference corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 215–226, Cham. Springer International Publishing.

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources*

and Evaluation (LREC'16), pages 1579–1585, Portorož, Slovenia. European Language Resources Association (ELRA).

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 650–659, Honolulu, Hawaii. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtskỳ, Anna Nedoluzhko, Michal Novák, and Daniel Zeman. 2021. Do UD trees match mention spans in coreference annotations? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3570–3576. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.

Marta Recasens and M Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language resources and evaluation*, 44(4):315–345.

Ina Roesiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ina Rösiger. 2018. BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 789–795, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Svetlana Toldova and Max Ionov. 2017. Coreference resolution for Russian: the impact of semantic features. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"*, pages 339–348.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26(1):95–128.

Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017. Joint UD parsing of Norwegian Bokmål and Nynorsk. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland.

Veronika Vincze, Klára Hegedűs, Alex Sliz-Nagy, and Richárd Farkas. 2018. SzegedKoref: A Hungarian coreference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*, 3(3):3–4.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Voldemaras Žitkus and Rita Butkiene. 2018. Coreference annotation scheme and corpus for Lithuanian language. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 243–250. IEEE.