data

MDPI

Article

Neural Coreference Resolution for Dutch Parliamentary Documents with the DutchParliament Dataset

Ruben van Heusden [1,*], Jaap Kamps [2] and Maarten Marx [1,*]

[1] Information Retrieval Lab, University of Amsterdam, 1098 XH Amsterdam, The Netherlands
[2] Faculty of Humanities, University of Amsterdam, 1012 GC Amsterdam, The Netherlands
* Correspondence: r.j.vanheusden@uva.nl (R.v.H.); maartenmarx@uva.nl (M.M.)

Abstract: The task of coreference resolution concerns the clustering of words and phrases referring to the same entity in text, either in the same document or across multiple documents. The task is challenging, as it concerns elements of named entity recognition and reading comprehension, as well as others. In this paper, we introduce DutchParliament, a new Dutch coreference resolution dataset obtained through the manual annotation of 74 government debates, expanded with a domain-specific class. In contrast to existing datasets, which are often composed of news articles, blogs or other documents, the debates in DutchParliament are transcriptions of speech, and therefore offer a unique structure and way of referencing compared to other datasets. By constructing and releasing this dataset, we hope to facilitate the research on coreference resolution in niche domains, with different characteristics than traditional datasets. The DutchParliament dataset was compared to SoNaR-1 and RiddleCoref, two other existing Dutch coreference resolution corpora, to highlight its particularities and differences from existing datasets. Furthermore, two coreference resolution models for Dutch, the rule-based DutchCoref model and the neural e2eDutch model, were evaluated on the DutchParliament dataset to examine their performance on the DutchParliament dataset. It was found that the characteristics of the DutchParliament dataset are quite different from that of the other two datasets, although the performance of the e2eDutch model does not seem to be significantly affected by this. Furthermore, experiments were conducted by utilizing the metadata present in the DutchParliament corpus to improve the performance of the e2eDutch model. The results indicate that the addition of available metadata about speakers has a beneficial effect on the performance of the model, although the addition of the gender of speakers seems to have a limited effect.

Keywords: information retrieval; coreference resolution; datasets

Citation: van Heusden, R.; Kamps, J.; Marx, M. Neural Coreference Resolution for Dutch Parliamentary Documents with the DutchParliament Dataset. *Data* **2023**, *8*, 34. https://doi.org/10.3390/data8020034

Academic Editor: Henning Müller

Received: 28 December 2022
Revised: 21 January 2023
Accepted: 30 January 2023
Published: 1 February 2023

1. Introduction

Coreference resolution can be defined as the task of clustering mentions of named entities in text into coherent clusters referring to the same entity. In Example 1, an example of coreference resolution on a Dutch sentence is shown.

Example 1. *"Ik heb op **Mark Rutte** gestemd, omdat ik **zijn** visie op de toekomst van het land deel", zei Emma*

In Example 1, two entities are present—Mark Rutte and Emma—and both have several references to them, indicated by boldface and underlined text. In this particular example, both the references to Mark Rutte and Emma are by the use of a pronoun, but the references themselves can also be nouns or noun phrases (for example, de heer Rutte).

Although early coreference resolution systems were almost exclusively rule-based and revolved around hand-crafted ruled and parse trees [1–3], most of the recent approaches to coreference resolution have involved a form of supervised machine learning or statistical learning [4–7]. This shift in the approach to coreference resolution has increased the need for

annotated training data, more than for the rule-based models. Most of the previous work on creating datasets for coreference resolution has focused on the English language, and most of the datasets are on genres such as news articles or Internet blogs [8–10]. To obtain a good generalization performance of these methods on more niche genres, it is important that datasets on more nice domains are created so as to not bias models towards 'easy' examples or samples that are all from the same domain, something that will most likely result in a poor generalization performance of the models. Previous work on coreference resolution for the Dutch language has mostly focused on documents originating from news articles, such as in the SoNaR-1 corpus, and Shared Task 1 of the SemEval 2010 challenge [9,11], as well as recent work on coreference resolution in Dutch literary text [12]. The DutchParliament dataset presented in this research is focused on applying coreference resolution to a new genre of documents, namely parliamentary proceedings. This type of document provides a unique challenge to the existing models, as the documents are of a very different nature than most documents currently used in coreference resolution, exhibiting much more of a resemblance to conversational text. The difference in the contents and nature of the DutchParliament dataset makes it a useful addition to the existing datasets, facilitating further research on coreference resolution for specific domains. The main contributions of this research are as follows:

- We present a new coreference resolution dataset consisting of Dutch parliamentary documents annotated with coreference resolution links and a rich set of metadata features, called DutchParliament.
- We performed a comparative analysis of the DutchParliament corpus with two other Dutch coreference resolution datasets, namely the SoNaR-1 corpus and the RiddleCoref corpus. We investigated the overall structure and size of the corpus, and compare various lexical statistics.
- We evaluated two existing models for Dutch coreference resolution on the DutchParliament dataset and discuss their performance in comparison with the SoNaR-1 and RiddleCoref datasets.
- We conducted several experiments regarding the addition of metadata to the e2eDutch model and found that, for the parliamentary meetings, the addition of metadata about the speaker of utterances has a substantial positive effect on the performance of the model. The addition of the metadata about the gender of speakers does not seem to have any significant effect on the model.

2. Related Work

One of the earliest algorithms developed for the task of coreference resolution is Hobbs' algorithm [1], introduced in the 1970s. This algorithm relies solely on syntactic parse trees of sentences, and consists of a set of handcrafted rules to select coreference chains. Later work based on Hobbs' algorithm has also made use of syntactic parse trees, combined with more heuristic information or additional mechanisms [2,3]. An early method for the task of pronominal anaphora resolution was the Lappin and Leass algorithm, which introduced the concept of the salience assignment principle [3]. This principle assigned scores to all possible antecedents of a given pronoun, based on features such as the distance between the antecedent and the pronoun and the context in which an NP occurs (not contained in another NP, being the subject in a sentence).

One of the early examples of using discourse information for pronominal coreference resolution is the BFP algorithm [13]. It performed this by using centering theory, which assumes certain rules and patterns followed in normal discourse based on the subject of the discourse. Centering theory revolves around Forward Looking Centers and Backward Looking Centers, where backward looking centers are the current entity that is the focus of the discourse, and forward looking centers are entities that could possibly be the new focus of the discourse. Based on the forward and backward looking centers, and the transitions (keep the same entity as the center or shift), coreference resolution was performed by ranking entities with several rules and preferring entities where the center was not shifted,

as this is most common in natural discourse. Later, several additions and changes to this model were proposed [2,14,15].

An early attempt at the incorporation of metadata regarding pronominal anaphora resolution was presented in [16]. Here, metadata were included by using WordNet and name lists to determine the gender for names (which are detected by the usage of capital letters) and added this attribute to the possible antecedents. As well as incorporating gender information, information on the *animacy* was also added. Animacy indicates whether an entity can be referred to by gender pronouns (he, she, her, his, etc.). This information is also extracted from WordNet and added as an attribute to the antecedents.

The *e2e architecture* was proposed in a paper by [7] and is one of the first examples of an end-to-end coreference resolution model. The e2e architecture is based on the long short-term memory (LSTM) network, and attempts to substitute the pipeline approach used in much of the previous research for an end-to-end approach using a neural network and word embeddings. In their research, it was found that the e2e model is successful in outperforming the baselines of that time, achieving a state-of-the-art performance on the CoNLL-2012 shared task dataset.

For the task of coreference resolution for the Dutch language, several approaches have been developed, with the two most recent approaches being the rule-based DutchCoref model [12] based on parse trees and handwritten rules, and the e2eDutch model [17], based on the aforementioned e2e model. The e2eDutch model is largely similar to the original e2e model, but replaces static word embeddings with word embeddings from BERT. Earlier methods of coreference resolution for Dutch were largely rule-based, and based on the idea pioneered in Soon et al. [18], where coreference resolution was performed by constructing vector representations for words using syntactic and semantic features. These features are then fed into a *mention-pair* algorithm, which outputs pairwise decisions that can be transformed into coreference cluster predictions. For the Dutch language, the work by Kobdani and Schütze [19] is largely based on the aforementioned approach.

As previously mentioned, prior work has been carried out on coreference resolution for the Dutch language. An example of this can be found in [20], with the construction of the COREA corpus and the evaluation of a coreference resolution model for Dutch on the created corpus, with the aim of improving the performance on a question-answering task. The COREA corpus is constructed mostly of Dutch news articles from the DCOI project and spoken text from the Corpus Gesproken Nederlands (CGN).

Another coreference resolution dataset was developed during the Sonar project, in which text from various sources, such as news articles, Wikipedia articles and social media posts from Twitter, was collected. This resulted in a corpus of 500 million tokens, called SoNaR-500 [9]. Of these 500 million tokens, a subset of 1 million tokens was annotated with more detailed semantic information, including coreference resolution annotations and named entity labels (the SoNaR-1 corpus). Although the annotations were produced automatically, they were manually corrected in the case of the SoNaR-1 corpus.

An example of a dataset that departs from the default type of data is the RiddleCoref corpus developed by [21], which concerns excerpts from Dutch novels. This dataset contains 33 documents, which are all extracts from either Dutch novels or Dutch translations of novels. The RiddleCoref is of particular interest to the current research, as it also concerns a genre of text that is different from datasets such as SoNaR-1, namely novels.

3. Materials and Methods

The DutchParliament dataset consists of parliamentary documents from the official XML records of parliamentary meetings of the Dutch government, collected during the ParlaMint project [22]. The dataset consists of debates held during meetings of the Dutch Lower House and Upper House, and the documents are transcriptions of those sessions. From this original dataset, 74 documents specifically concerning question sessions were selected for manual coreference annotation and then annotated by two human annotators. Question sessions were chosen specifically because the general structure resembles that of

spoken text, with interactions between participants, and thus provides a unique type of data not commonly found in Dutch datasets. The complete DutchParliament dataset can be downloaded from the DANS Easy data repository. [1]

3.1. Annotation Scheme

The annotation scheme that was used during the annotation process is an adaptation of the scheme used by [23] for Dutch coreference resolution, with the definitions of what types of spans and words are eligible for annotation being almost identical to the criteria used in [23]. One of the challenges that we faced during the data collection and annotation was the difficulty in annotating complex nested coreference resolution, in combination with a multitude of speakers in any given document. To this end, we decided to restrict the types of entities that are considered during annotation, as well as to exclude nested entities and include information on the speaker of a piece of text for each sentence. Four types of entities were considered for annotation, namely PERSON, ORG, LOC and MISC. As with previous works, the annotation of certain linguistic phenomena such as the pleonastic *it* (*It was raining*) was prohibited. In the case of composite named entities such as 'Nederlandse burgers', the guidelines from [23] were followed, and the entity was annotated as miscellaneous.

Apart from the four entity types mentioned, a new class of entity was added for this particular dataset, concerning various types of legal and political document references, which will be referred to as the official documents class.

Official Documents Class

The *Official Documents Class* contains a variety of mentions of official documents, ranging from laws and treaties to motions and letters from the parliament. The class is different from the LAW type used in named entity recognition, as it also includes letters and documents that might not be official laws, but that are relevant in parliamentary debates, such as letters sent or received by ministers. Table 1 shows several examples of entity clusters concerning laws and letters.

Table 1. Examples of clusters of the *official document* type.

Cluster	Cluster Words
Wet op kinderopvang	de Wet op de kinderopvangtoeslag, die regeling, de Wet kinderopvang, die wet
De woningwet	de Niewe Woningwet, de wet, die wet
Brief	brief aan de kamer, deze brief, die brief

3.2. Annotation Process

For the annotation of the coreference resolution clusters, the CorefAnnotator Tool [24] was used. Documents were converted from the XML format to plain text, and each sentence was preceded with the person that uttered the sentence to avoid ambiguity in the annotation process and provide the annotators with the metadata present in the original XML files. These indicators were removed after the annotation process. After the annotators had annotated the files, the annotations were converted from the *xmi.gz* format used by the CorefAnnotator tool to the CoNNL[2] format commonly used for coreference resolution systems.

Inter Annotator Agreement

For the calculation of the IAA scores described below, a separate set of 5 documents was annotated by both annotators to measure the agreement score.

To measure the inter-annotator agreement between the two annotators, the built-in score from the CorefAnnotator tool was used to calculate a rough agreement on the mention

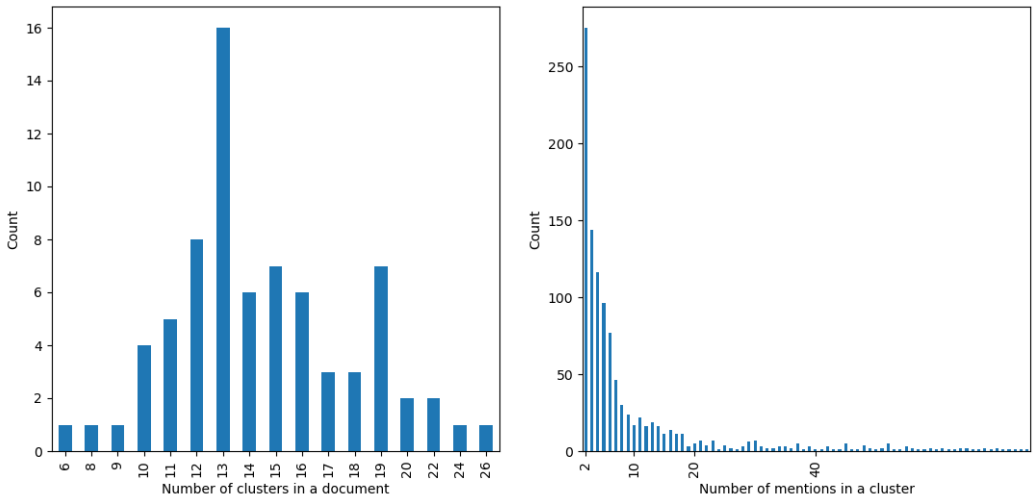
annotation and evaluate the agreement of the annotators, both during the annotation process as well as for the IAA evaluation. The agreement operates on mentions, and calculates the Jaccard similarity between the sets of annotated mentions of both annotators. This implementation considers annotations to be equal when the spans of the mentions are equal for both annotators, regardless of the entity to which they refer, and does not consider entity links in the evaluation. The average agreement score between the annotators was 77.9%, which is quite a high agreement score. For a more trustworthy evaluation of the mention annotation agreement, Cohen's kappa was also calculated between the two annotators, which resulted in an agreement score of roughly 88%

Although the mention agreement is relatively high for both mention annotation scores, correctly annotating mentions is only part of coreference resolution, and the correct linking of the mentions is equally if not more important. To measure the annotator agreement in links between mentions, the approach described in [25] was used, where the MUC score is used as a measurement of IAA on coreference links, where the annotations of one annotator are taken as gold standard, and the annotations of the other are taken as system output files. Calculating the IAA score for coreference links yielded a MUC F1 score of 86.62%, again indicating a high agreement between the two annotators. For the Jaccard similarity, singleton mentions were also included in the agreement evaluation, and excluded from MUC, as it is a linked-based metric: as singletons have no links, they have no effect on the score.

3.3. Dataset Statistics

The DutchParliament dataset contains 74 documents, with a total of roughly 180,000 tokens and 1753 coreference clusters. Each document contains, on average, 15 clusters, and those clusters, on average, have 7 mentions.

Figure 1a shows the distribution of the number of clusters in a file, with an average of 15 clusters in a file. It can be seen that the number of clusters in a file somewhat resembles a normal distribution, with files containing a minimum of 6 and a maximum of 26 clusters.



(**a**) (**b**)

Figure 1. Distributions of mentions and clusters for the DutchParliament dataset. (**a**) Distribution of the number of clusters in the 74 documents, with the number of clusters in a file on the x-axis and the number of occurrences on the y-axis. (**b**) Distribution of the number of mentions in a cluster for the 74 documents, with the size of the cluster on the x-axis and the number of occurrences on the y-axis.

Figure 1b shows the distribution of the number of mentions in a cluster, with the vast majority of the clusters in the dataset having 6 or fewer mentions, with a relatively long tail. On average, a coreference cluster contains roughly 7 mentions. Upon investigation, it was found that the larger clusters often concerned a minister, which is unsurprising given the fact that they were often the subject of the question sessions.

Another aspect of the dataset that can be examined is the distribution of the part-of-speech tags of all mentions in the dataset. For this, the mentions in all clusters were automatically tagged with a Dutch part-of-speech tagger from the SpaCy NLP package [26].

Accounting for some errors in the part-of-speech tagger, pronouns, nouns and proper nouns are by far the most common part-of-speech tags, as shown in Figure 2. This is probably due to the fact that the debate documents are transcripts of spoken text, in which these kind of pronominal references are usually much more common than in text that was originally written, as pronominal references tend to be relatively common in spoken texts.

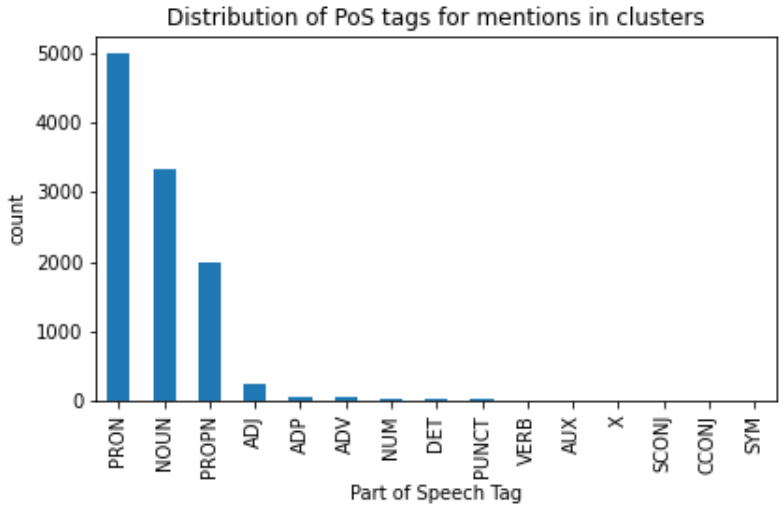


Figure 2. Part-of-speech tag distribution for all mentions in the dataset.

For a similar reason, the high percentage of mentions being a noun most likely also originates from the specific usage of language in these parliamentary documents. When examining the most often occurring mentions in the NOUN part-of-speech category, it wsa found that words such as 'de minister', 'de staatssecretaris' or 'kamerlid' were used frequently, thus accounting for a substantial amount of noun references in the dataset.

In addition to the part-of-speech tags, the distribution of the different types of named entities represented by the coreference clusters was also investigated. As the named entity tags present in the original XML files might not correspond to tokens selected by the annotators, these named entity tags were not used. Instead, a Dutch named entity classifier from SpaCy [26] was used to classify every mention in the dataset. This results in clusters where each mention has an entity type, which might not be consistent for all mentions in a cluster. In this case, the majority vote was taken to obtain the entity type of the entire cluster.

Figure 3 shows the distribution of named entity types represented by the coreference clusters. Here, each count in one category signifies a cluster of that type. For example, a cluster containing 'de minister president', 'hij', 'de heer Rutte' and 'Mark Rutte' would be counted as a PERSON entity cluster. As can be seen in Figure 3, PERSON type is the most frequently occurring entity type in the dataset, accounting for nearly 50 percent of all clusters. Upon further investigation of the PERSON clusters, it was found that, for the PERSON category, the pronoun speech tag is the most occurring part-of-speech tag, accounting for 60% of the tokens. Together with the noun category (30%) and the proper noun category (7%), these three categories make up almost all part-of speech-tags in the PERSON entity clusters.

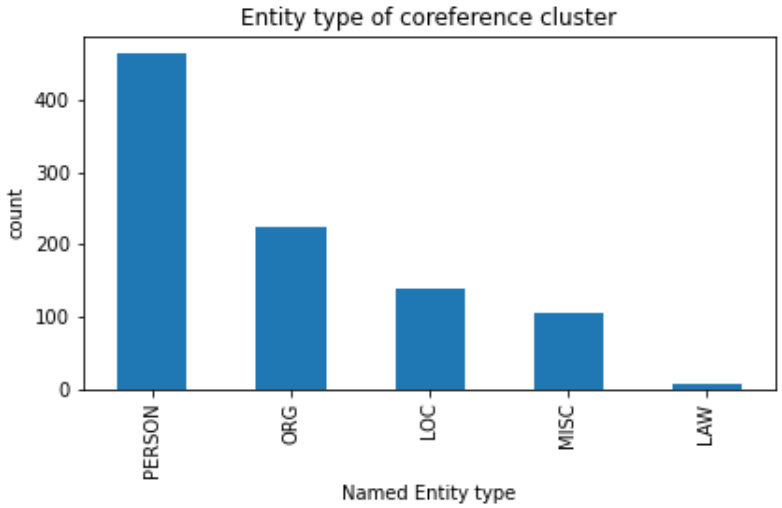


Figure 3. Distribution of named entity categories in the parliamentary debates.

4. Dataset Comparison

In this section, the DutchParliament dataset will be compared to two existing Dutch datasets, namely RiddleCoref and SoNaR-1. The main goal of this comparison is to investigate the difference in characteristics between more classic datasets and the more conversation-like DutchParliament dataset, and whether or not these differences impact the performance of coreference resolution models. The datasets were compared on general statistics, lexical statistics and the performance of two models for Dutch coreference resolution.

4.1. General Structure Analysis

For the general statistics of the DutchParliament corpus, the data were gathered from the raw CoNLL files and aggregated for the entire corpus. The dataset statistics of the SoNaR-1 corpus and the RiddleCoref corpus were partly taken from the paper of [21], where the statistics from development and test sets were aggregated, and the statistics that were not present in [21] were calculated manually.

One of the first observations that can be made from Table 2 is the relatively low number of clusters and entity mentions in the DutchParliament dataset in comparison to the other two datasets. This is mostly due to the restricted annotation scheme used for the construction of the DutchParliament dataset, which excluded many coreference mentions that are present in the other two corpora. The RiddleCoref dataset and the DutchParliament dataset are quite comparable in size when comparing the number of tokens, with the SoNaR-1 corpus being significantly larger in size than the other two corpora. When examining the average lengths of the documents, the average document length in DutchParliament is almost two times shorter than the average document length in the RiddleCoref corpus, but significantly longer than the average document length in the SoNaR-1 corpus. This is most likely due to the different genres of the corpora, with news articles often being relatively short in length, and the novel extracts being relatively long. It can also be noted that the average sentence length appears to be very similar between the three corpora, ranging between 16.3 and 17.6 tokens, suggesting that at least the sentence structure of the three corpora is relatively similar.

Table 2. Comparison of various dataset statistics for SoNar-1, RiddleCoref and DutchParliament. For the statistics of the RiddleCoref corpus, the numbers were taken directly from [21].

	SoNaR-1	**RiddleCoref**	**DutchParliament**
Number of files	862	33	74
Number of tokens	approx. 1,000,000	approx. 160,000	approx. 180,000
Number of sentences	59,602	9864	11,038
Average sentence length in tokens	16.6	17.6	16.3
Number of clusters	205,103	14,692	1082
Number of mentions	289,955	38,647	10,771
Fraction of coreference tokens	0.29	0.24	0.60
Average document length in tokens	1160	4897	2432
Average document length in sentences	69.1	298.9	149.1

4.2. Lexical Statistics Comparison

Apart from a study of the general statistics of the datasets, the lexical similarity between datasets was also investigated. In this case, the evaluation was only conducted between the SoNaR-1 and the DutchParliament corpus because of limited access to the RiddleCoref corpus. For this comparison, the number of unique tokens, number of tokens divided by the number of unique tokens, the lexical frequency profile (LFP) [27] and Yule's K were reported, similar to the research presented in [28]. The LFP consists of multiple *frequency bands* of word frequencies that were calculated. Similar to [28], three frequency bands were used, namely: (1) the percentage of tokens in the corpus occurring in the 1000 most frequent words of a language (B1); (2) the percentage of tokens in the corpus occurring in the next 1000 most frequent words of a language (B2); (3) the percentage of the rest of the tokens not occurring in the top 2000 (B3). For the Dutch word frequency list, we used the SUBTLEX-NL word list [29].

The Yule's K metric is a measure of the number of unique words in the text and provides an estimate on the on the variation in words and their frequencies in the corpus. For Yule's K, the bigger the value, the less diverse the corpus vocabulary.

From Table 3, it can be seen that, although the number of unique tokens present in the SoNaR-1 corpus is much higher than the number of tokens present in the DutchParliament dataset, the token type ratio and Yule's K are relatively similar. Another interesting observation that can be made is that, for the DutchParliament corpus, the B1 value is much higher than the B1 value for the SoNaR-1 corpus. This indicates that the DutchParliament corpus consists of more words that occur very frequently in Dutch in general. A possible explanation for this could be that the DutchParliament corpus contains relatively more pronouns than the SoNaR-1 corpus and that these pronouns are in the most frequently occurring words in the Dutch word frequency list. The dataset genre, conversational text, might also play an important role in this.

Table 3. Comparison of DutchParliament and SoNaR-1 and the number of unique tokens, token/type ratio, three LFP frequency bands and Yule's K.

Corpus	**Number of Unique Tokens**	**Token/Type Ratio**	**B1**	**B2**	**B3**	**Yule's K**
DutchParliament	10,749	16.313	0.73	0.05	0.22	99.86
SoNaR-1	65,632	13.827	0.58	0.06	0.37	102.19

4.3. Model Performance Comparison

In this section, the performance of two coreference resolution models for the Dutch language is compared across the DutchParliament, RiddleCoref and Sonar-1 corpora to investigate the effect of the difference in genres and structure of these corpora on the perfor-

mance of these models. For the DutchParliament dataset, the models were retrained (in the case of the e2eDutch model) and run on the dataset. For the RiddleCoref and SoNaR-1 corpora, the performances of the models were taken directly from [21], where the scores that excluded singletons were taken, to ensure a fair comparison.

4.3.1. DutchCoref Rule-Based System

The DutchCoref system is a rule-based coreference resolution system utilizing Alpino [30] parse trees in combination with a set of rules for determining coreference clusters in text [12]. The code for this rule-based model is publicly available.[3] Before running the coreference resolution system, the debates were parsed with the Alpino parser. In order to accommodate for the altered annotation guidelines used for the DutchParliament corpus, the model was set to follow the SoNaR-1 guidelines, for which, the annotation scheme was somewhat more similar to the scheme used for DutchParliament compared to the default setting. For correct parsing by the Alpino parser and the subsequent coreference resolution system, each speaker turn was separated by a new line in the text files of the corpus.

4.3.2. e2eDutch

The e2eDutch model is an adaptation of the e2e model by [7], adapted for the Dutch language, based on a bidirectional LSTM model that uses contextual BERT embeddings as the input. The model combines mention detection with the linking of coreference mentions using a pairwise mention classifier. For the DutchParliament dataset, a new model was trained on the dataset. The model was trained with the default hyper parameters of five epochs and a learning rate of 1×10^{-4}, and embeddings from the Dutch BERT model *Bertje*[4] were used.

4.3.3. Evaluation

There exist a multitude of coreference resolution metrics, each of them suited for slightly different applications and requirements of a coreference resolution system, such as a focus an mention detection or mention linking. In this research, two separate metrics were used to evaluate the coreference resolution systems, namely the LEA score and the mention score. The LEA metric is a link-based metric introduced in Moosavi and Strube [31] that is based on calculating the quality of coreference resolution through the combination of the quality of the resolution of an entity as well as its importance in the overall prediction. It overcomes some of the shortcomings that earlier evaluation metrics had by weighting entities by importance while also considering coreference links, and, as such, has become an often-used metric in coreference resolution.

As taken from [31], the *LEA* score is calculated as given in Equation (1)

$$LEA = \frac{\sum_{e_i \in E} (importance(e_i) \times resolution\_score(e_i))}{\sum_{e_k \in E} importance(e_k)} \quad (1)$$

where e_i is an entity in the gold standard set. In this equation, the importance is usually taken as the size of the entity, i.e., $|e_i|$, but other measures can also be used depending on the specific task. The *resolution_score* is similar to other link-based metrics, such as the BLANC score, and consists of the following:

$$resolution\_score(k_i) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)} \quad (2)$$

where k_i is a coreference cluster from the gold standard set, and r_i is an entity in the system output. The $link()$ function counts the number of links in an entity cluster. When Equation (2) is plugged into Equation (1), this will yield the recall of the system. To obtain the precision score of the model output, the same formulas are used; however, the roles of the gold standard set and the system output set are reversed.

The mention score is another score that can be used to evaluate coreference resolution systems by comparing the mentions in the prediction and the gold standard. The mention score is important for evaluation, as mentions are at the core of the coreference resolution task, and missing mentions or producing erroneous mentions can have a large impact on the overall performance of the model. The formula for the mention score is given below:

$$mention\_score = \frac{|K \cap R|}{|K|} \tag{3}$$

Here, K is the set of mentions in the gold standard set, and R is the set of mentions in the output generated by the system. As with the LEA metric, Equation (3) calculates the recall of the system, and, to calculate the precision, the roles of the gold standard and system output sets have to be reversed, with the system output acting as K and the gold standard set acting as R. Note that, in this version of the metric, mentions have to be completely equal to be counted as correct.

For the evaluation of the coreference resolution systems, the dataset was split up into a train set, a development set and a test set, consisting of 60%, 20% and 20% of the dataset, respectively. As mentioned previously, the DutchCoref baseline model is a rule-based model; thus, the behaviour of the model cannot be altered through training. For this reason, the performance of the model was only evaluated on the test set in order to conduct a fair comparison. For the DutchParliament dataset, all singleton mentions were removed from the dataset. The code used to run the experiments in this paper is available on GitHub. [5]

One of the first observations that can be made from viewing the results presented in Table 4 is the large gap between the performance of the rule-based DutchCoref model and the neural e2eDutch model on the DutchParliament dataset for both the LEA and the mention score. When examining the recall and precision separately, it was found that the precision was 25.0 and the recall was 38.5, showing that this decreased performance is mostly due to a decrease in precision.

Table 4. Results of the DutchCoref and e2eDutch models on the DutchParliament, RiddleCoref and SoNaR-1 datasets. The scores for the SoNaR-1 and RiddleCoref datasets were taken from [21] (all reported scores are with singletons excluded from evaluation).

Model	**DutchParliament**		**RiddleCoref**		**SoNaR-1**	
	Mentions F1	**LEA F1**	**Mentions F1**	**LEA F1**	**Mentions F1**	**LEA F1**
DutchCoref	46.99	32.00	**80.56**	**48.15**	63.57	39.71
e2eDutch	**76.57**	**45.54**	79.94	45.31	**67.08**	**46.18**

The most probable explanation for this low precision behaviour is the difference in annotation scheme used in the DutchParliament dataset. The DutchCoref model was developed with a different annotation scheme in mind, and, as such, it considers a much broader set of words and phrases as candidates for coreference clusters. As the scheme used in the DutchParliament dataset is much more restricted, the DutchCoref model will 'overclassify' and annotate many more mentions as co-references than the scheme considers. As a result, the recall of the model is relatively high, while comprising the precision of the model. This is confirmed by looking at the average number of clusters in the parliamentary debates as predicted by the DutchCoref model. In the gold standard set, the average number of clusters was approximately 15, with the average number of clusters in the output of the DutchCoref model being approximately 20, showing a significant increase in the number of predicted clusters.

Another observation that can be made is that the performance of the e2eDutch model on the DutchParliament dataset is quite close to that of the performance of the model on the RiddleCoref dataset. When looking at the LEA metric, the performance of the e2eDutch model on the DutchParliament dataset is very similar; however, the F1 score from the

mentions is slightly lower, suggesting that the detection of the mentions is somewhat more problematic for the DutchParliament dataset. When comparing the performance of the e2eDutch model on the DutchParliament and SoNaR-1 corpus, it can be seen that the scores of the e2eDutch model are again very close between the two when observing the LEA metric. The mention score of the e2eDutch model is much higher on the DutchParliament dataset when compared to the SoNaR-1 dataset.

Based on the results discussed above, it seems that, whereas the DutchCoref model performed worse on the DutchParliament dataset, the e2eDutch model was not heavily affected by the different nature of the parliamentary documents, and that, despite the differing characteristics of the DutchParliament dataset and the high number of pronouns, the model obtained similar LEA F1 scores across all three datasets.

5. Metadata Addition

In this section, the results of adding various types of metadata from the DutchParliament corpus to the e2eDutch model are presented.

5.1. Inclusion of Speaker Metadata

The original e2e model for English has the option of adding metadata indicating the speaker of a piece of text to the input of the model. To achieve this, each speaker is assigned a unique ID, and all words uttered by that speaker are assigned the ID of that speaker. During the training of the e2e model, an embedding is learned for each speaker, which is then appended to each candidate mention and used in the pairwise comparison of mentions. To use this information in the DutchParliament dataset, the speaker information was extracted from the XML files, and each word was assigned the appropriate speaker ID. Table 5 shows an example of the speaker metadata used in the e2e model.

Table 5. Illustration of the assignment of unique IDs to words based on the speaker of the sentence. Unique IDs are calculated per file and thus IDs for a speaker can be different in different files.

Geert Wilders	**1**	**1**	**1**	**1**
Sentence	Doe	eens	normaal	man
Mark Rutte	2	2	2	
Sentence	Doe	zelf	normaal	

5.2. Inclusion of Gender Metadata

Although the e2e model has the capability of handling metadata about the speaker of a sentence, it does not have an option for adding the gender of a speaker to information used in the pairwise comparison of candidate mentions. It might, however, be beneficial to add this type of metadata in order to correct possible mistakes such as 'zijn', and 'haar' ending up in the same cluster. However, adding this type of information into the e2e model is not straightforward, as this type of information can only be encoded on the level of the speaker, making it more difficult to assign the gender information to individual tokens. To alleviate this problem, another, much simpler approach was used. In this approach, tokens were assigned either a 0, 1 or 2 based on the 'gender' of the word. In this case, only pronouns and names were actually assigned a nonzero number, as these are the tokens of interest, and all of the other tokens received a zero. For the pronouns, a handmade list of male and female pronouns was constructed, and, for names, a list of common Dutch names and their most commonly associated gender was used. The number assigned to the token was converted to a one-hot vector that was appended to the vector of the token for pairwise comparison.

5.3. Results

When comparing the 'plain' version of the model with the versions of the model with different types of metadata added in Table 6, improvements in the performance can be

observed. The performance improvement seems to be most prominent for the version of the model for which only the speaker metadata are added.

Table 6. LEA and mention scores of the e2eDutch model with different types of metadata added to the model on the DutchParliament dataset.

Model	LEA			Mentions
	Precision	Recall	F1	F1
e2e-Dutch	52.06	41.57	45.54	76.57
e2e Dutch + speakers	54.59	**50.60**	**51.95**	**77.83**
e2e Dutch + gender	51.20	45.20	47.48	75.44
e2e Dutch + both	**58.35**	42.24	47.71	74.85

With only the addition of the speaker metadata, the recall score of the e2eDutch model increases substantially in comparison with the plain e2eDutch model. The most likely explanation for the improved recall of the model is that the metadata help to break up clusters that contain multiple persons, as these person clusters are most likely affected the most by the introduction of the speaker metadata. For example, consider words such as 'ik' and 'mijn'. Without information about the speaker, the model assumes that multiple occurrences of these words all refer to the same cluster, as their surface forms are identical. However, different people could refer to themselves through words such as 'ik' and 'mijn', and thus correctly classifying these words in different clusters can lead to an improved recall score of the model. This can be confirmed by looking at the average number of clusters in the files, which increased from roughly 8.5 to 10 when metadata were added. This suggests that the addition of the speaker metadata indeed splits up the clusters more accurately.

However, when both of these types of metadata are combined, it can be observed that there is a substantial increase in the precision of the e2eDutch model when compared to the plain version of the model, albeit at the cost of a decreased recall performance when compared with only adding speaker or gender metadata. Although the exact reason for this behavior is not completely clear, one possibility is that the model uses the gender information in combination with the speaker information to 'remove' conflicting mentions from a cluster, such as both 'zijn' and 'haar' in a cluster.

When examining Figures 4 and 5, it can be seen that the distribution of part-of-speech tags is significantly altered when adding both the gender and speaker metadata, with the percentage of pronouns being quite a bit lower. This would suggest that the addition of the gender and speaker metadata simultaneously is indeed distributing the pronouns into clusters more fairly, lowering the average percentage of pronouns in a cluster. However, the model might end up clustering mentions based mostly on their gender, creating fewer, larger clusters partitioned based on the gender of mentions, explaining the decreased recall.

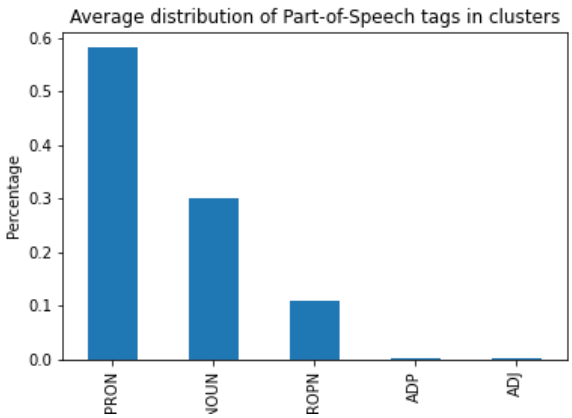


Figure 4. Average distribution of part-of-speech tags in coreference clusters without the addition of any metadata.

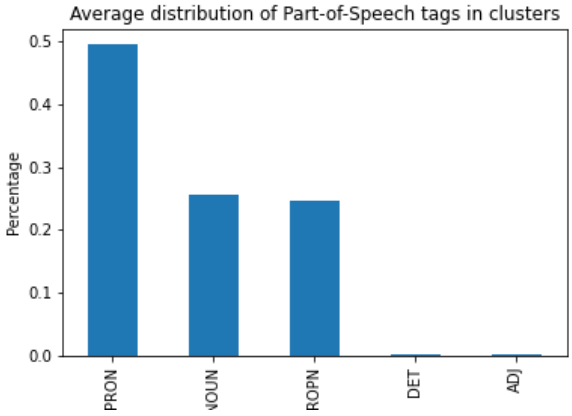


Figure 5. Average distribution of part-of-speech tags in coreference clusters with using both gender and speaker metadata.

5.4. Error Analysis

To gain more insight into the performance of the model and provide some examples of sentences in the dataset itself, a small error analysis was performed here by showing several examples of correct entity coreference links and several example of model mistakes.

Example 2. *Examples of correct and incorrect entity coreference for e2e-Dutch without metadata*

a. *(correct) Daar verheug* $[ik]_1$ $[mij]_1$ *ook op , maar dat is echt een ander verhaal .*

b. *(incorrect)* $[De\ minister]_1$ *zegt dat er veel lokale bedrijven worden ingezet , ook uit Drenthe en Friesland , en dat er zo ' n 2.500 vakkrachten , zzp'ers , worden ingehuurd en kunnen worden ingehuurd . Dat klinkt mooi , maar daarom was* $[ik]_1$ *zo verbaasd over het artikel en dan met name over het citaat dat* $[Centrum\ Veilig\ Wonen]_2$ *een soort monopoliepositie zou hebben doordat het alle touwtjes in handen heeft .*

c *(incorrect)* $[ik]_1$ *had een specifieke vraag gesteld over het gesprek dat* $[de\ premier]_2$ *en* $[de\ minister]_2$.

Example 3. *Examples of correct and incorrect entity coreference for e2e-Dutch wit speaker metadata*

a. *(correct) Is* $[de\ staatssecretaris]_1$ *bereid die uitzonderingsregel ook voor hen te overwegen? De regel rond het wettelijk doorwerkvereiste is natuurlijk iets voor* $[mijn]_1$ $[collega\ van\ Financiën]_3$. $[ik]_1$ *verzoek* $[de\ heer\ Krol]_2$ *dan ook om het bij* $[hem]_3$ *aan de orde te stellen.*

Example 2 shows several example sentences, annotated with coreference links with the e2e-Dutch model without any metadata. Example (2-a) shows an easy example of a coreference, where the two words that should be linked are right next to each other. (2-b) and (2-c) are more difficult, with (2-b) being incorrect as 'Minister', and 'Ik' incorrectly referring to the same entity. It is possible that the longer distance between the two words is a cause of this. Example (2-c) shows another peculiar error, where 'de premier' and 'de minister' are classified as the same entity, although they clearly do not belong to the same entity. An example of the speaker metadata proving useful can be seen in Example (3-a), where the first sentence is uttered by another speaker than the remainder of the text, and the speaker metadata can be used to distinguish 'de staatssecretaris' from 'Ik'.

In conclusion, the addition of speaker metadata to the e2eDutch model has a substantial positive effect on the performance of the model compared to the version of the model where these metadata are not present. This result highlights that the inclusion of metadata in coreference resolution for conversational text can be very beneficial, and that an effort should be made to include these data when possible. Although the addition of the gender metadata appears to have some positive effects on the precision of the model when combined with the speaker ID information, this addition alone does not have a large effect on the performance on the model.

6. Discussion and Future Work

In this research, a subset of parliamentary documents was selected for annotation, namely the question sessions from the Lower House. Although the dataset has a relatively large number of files, it might not be an accurate representation of parliamentary meetings in general, and it could be the case that using more parliamentary data would provide a more accurate representation of parliamentary documents. Regarding the addition of gender metadata, the current method of adding the gender metadata is quite rudimentary, and this simple approach might not be as effective in incorporating the information into the e2e model as methods that are more sophisticated. In future work, more research can be conducted into investigating the addition of metadata from structured files into neural models and the e2e model in particular, as it has the potential to provide the algorithms with more information about a variety of aspects of the text, such as the genders of the speakers or the general topic and structure of the texts.

7. Conclusions

In this research, a new Dutch dataset for coreference resolution, DutchParliament, was created, consisting of parliamentary debates from the Dutch government. A comparison was made between the developed corpus and the RiddleCoref and SoNaR-1 corpora, two other Dutch coreference resolution corpora. It was found that the DutchParliament dataset was significantly different from the SoNaR-1 and RiddleCoref corpora when comparing the corpora on general statistics and lexical similarity measures. In addition, two existing models for coreference resolution for Dutch were evaluated on the newly constructed dataset, where it was found that the e2e model appeared to achieve similar scores for the DutchParliament dataset compared to the other two datasets, despite the different genre of the DutchParliament dataset. Finally, the addition of both speaker and gender metadata present in the DutchParliament corpus to the e2eDutch model was evaluated. It was found that, although the addition of speaker metadata proved to be beneficial to the performance of the model, the addition of the gender metadata did not provide any substantial performance improvements.

Author Contributions: Conceptualization, M.M., J.K. and R.v.H.; resources, M.M. and R.v.H.; data curation, M.M. and R.v.H.; software, R.v.H. and M.M.; formal analysis, M.M., J.K. and R.v.H.; supervision, M.M. and J.K.; funding acquisition, J.K. and M.M.; validation, R.v.H. and M.M.; investigation, R.v.H. and M.M.; visualization, R.v.H. and M.M.; methodology, R.v.H., J.K. and M.M.; writing—original draft, M.M., J.K. and R.v.H.; project administration, R.v.H. and M.M.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Netherlands Organization for Scientific Research (NWO) through the ACCESS project grant CISC.CC.016.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both the data and the code used in this research are publicly available. The DutchParliament Dataset is located at DANS EASY [6] and the code is available on GitHub [7].

Conflicts of Interest: The authors have no conflicts of interest to declare that are relevant to the content of this article.

Notes

1 https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:224705/tab/2 (Accessed on 29 January 2023).
2 https://universaldependencies.org/format.html (Accessed on 29 January 2023).
3 https://github.com/andreasvc/dutchcoref (Accessed on 29 January 2023).
4 https://github.com/wietsedv/bertje (Accessed on 29 January 2023).
5 https://github.com/RubenvanHeusden/DutchParliamentCoreference/blob/main/README.md (Accessed on 29 January 2023).
6 https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:224705 (Accessed on 29 January 2023).

[7] https://github.com/RubenvanHeusden/DutchParliamentCoreference/blob/main/README.md (Accessed on 29 January 2023).

References

1. Hobbs, J.R. Resolving Pronoun References. *Lingua* **1978**, *44*, 311–338. [CrossRef]
2. Kameyama, M. A Property-Sharing Constraint in Centering. In Proceedings of the ACL'86: 24th Annual Meeting on Association for Computational Linguistics, New York, NY, USA, 10–13 July 1986; Association for Computational Linguistics: New York, NY, USA, 1986; pp. 200–206. [CrossRef]
3. Lappin, S.; Leass, H.J. An Algorithm for Pronominal Anaphora Resolution. *Comput. Linguist.* **1994**, *20*, 535–561.
4. Connolly, D.; Burger, J.D.; Day, D.S. A Machine Learning Approach to Anaphoric Reference. In Proceedings of the New Methods in Language Processing, Sydney, Australia, 11–17 January 1997; pp. 133–144.
5. Cardie, C.; Wagstaff, K. Noun Phrase Coreference as Clustering. In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, USA, 21–22 June 1999.
6. Ng, V. Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, USA, 25–30 June 2005; pp. 157–164.
7. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-end Neural Coreference Resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 188–197.
8. Pradhan, S.; Ramshaw, L.; Marcus, M.; Palmer, M.; Weischedel, R.; Xue, N. Conll-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Portland, OR, USA, 23–24 June 2011; pp. 1–27.
9. Oostdijk, N.; Reynaert, M.; Hoste, V.; Schuurman, I. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In *Essential Speech and Language Technology for Dutch*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 219–247.
10. Recasens, M.; Martí, M.A. AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Lang. Resour. Eval.* **2010**, *44*, 315–345. [CrossRef]
11. Recasens, M.; Màrquez, L.; Sapena, E.; Martí, M.A.; Taulé, M.; Hoste, V.; Poesio, M.; Versley, Y. Semeval-2010 Task 1: Coreference Resolution in Multiple Languages. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 1–8.
12. van Cranenburgh, A. A Dutch Coreference Resolution System with an Evaluation on Literary Fiction. *Comput. Linguist. Neth. J.* **2019**, *9*, 27–54.
13. Brennan, S.E.; Friedman, M.W.; Pollard, C. A Centering Approach to Pronouns. In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, Stanford CA, USA, 6–9 July 1987; pp. 155–162.
14. Strube, M.; Hahn, U. Functional Centering. *arXiv* **1996**, arXiv:cmp-lg/9605021.
15. Iida, R.; Inui, K.; Takamura, H.; Matsumoto, Y. Incorporating contextual cues in trainable models for coreference resolution. In Proceedings of the EACL Workshop on the Computational Treatment of Anaphora, Budapest, Hungary, 14 April 2003; pp. 23–30.
16. Liang, T.; Wu, D.S. Automatic Pronominal Anaphora Resolution in English Texts. *Int. J. Comput. Linguist. Chin. Lang. Process.* **2001**, *9*, 21–40.
17. van Kuppevelt, D.; Attema, J. e2e-Dutch. 2020. Available online: https://github.com/Filter-Bubble/e2e-Dutch (accessed on 29 January 2023). [CrossRef]
18. Soon, W.M.; Ng, H.T.; Lim, D.C.Y. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.* **2001**, *27*, 521–544. [CrossRef]
19. Kobdani, H.; Schütze, H. Sucre: A modular system for coreference resolution. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 92–95.
20. Hendrickx, I.; Bouma, G.; Coppens, F.; Daelemans, W.; Hoste, V.; Kloosterman, G.; Mineur, A.M.; Van Der Vloet, J.; Verschelde, J.L. A Coreference Corpus and Resolution System for Dutch. In Proceedings of the LREC, Citeseer, Marrakech, Morocco, 28–30 May 2008.
21. Poot, C.; van Cranenburgh, A. A Benchmark of Rule-Based and Neural Coreference Resolution in Dutch Novels and News. In Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference, Barcelona, Spain, 12 December 2020; Association for Computational Linguistics: Barcelona, Spain 2020; pp. 79–90.
22. Erjavec, T.; Ogrodniczuk, M.; Osenova, P.; Ljubešić, N.; Simov, K.; Grigorova, V.; Rudolf, M.; Pančur, A.; Kopp, M.; Barkarson, S.; et al. Linguistically Annotated Multilingual Comparable Corpora of Parliamentary Debates ParlaMint.ana 2.1, 2021. Slovenian Language Resource Repository CLARIN. SI. Online Resource. Available online: https://link.springer.com/article/10.1007/s10579-021-09574-0 (accessed on 29 January 2023). [CrossRef]
23. Schoen, A.; van Son, C.; van Erp, M.; van Vliet, H. *NewsReader Document-Level Annotation Guidelines-Dutch TechReport 2014-8*; Technical Report; VU University: Amsterdam, The Netherlands, 2014.
24. Reiter, N. CorefAnnotator—A New Annotation Tool for Entity References. In Proceedings of the Abstracts of EADH: Data in the Digital Humanities, Galway, Ireland, 7–9 December 2018. [CrossRef]
25. Hendrickx, I.; Hoste, V.; Daelemans, W. Semantic and Syntactic Features for Dutch Coreference Resolution. In *Lecture Notes in Computer Science, Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel, 17–23 February 2008*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 351–361.

26. Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. Online Resource. Available online: https://sentometrics-research.com/publication/72/ (accessed on 29 January 2023).
27. Laufer, B. The Lexical Profile of Second Language Writing: Does It Change Over Time? *RELC J.* **1994**, *25*, 21–33. [CrossRef]
28. Vanmassenhove, E.; Shterionov, D.; Gwilliam, M. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. *arXiv* **2021**, arXiv:2102.00287.
29. Keuleers, E.; Brysbaert, M.; New, B. SUBTLEX-NL: A New Measure for Dutch Word Frequency Based on Film Subtitles. *Behav. Res. Methods* **2010**, *42*, 643–650. [CrossRef] [PubMed]
30. Beek, L.V.D.; Bouma, G.; Malouf, R.; van Noord, G. The Alpino dependency treebank. In Proceedings of the Computational linguistics in the Netherlands, Twente, The Netherlands, 30 November 2001.
31. Moosavi, N.S.; Strube, M. 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, Berlin, Germany, 7–12 August 2016; pp. 632–642.