French coreference for spoken and written language

Rodrigo Wilkens[1], **Bruno Oberle**[1], **Frédéric Landragin**[2], **Amalia Todirascu**[1]
[1] LiLPa, [2] LaTTICe-CNRS
[1] University of Strasbourg, [2] École Normale Supérieure & Université Paris 3-Sorbonne Nouvelle
rswilkens@gmail.com, b.oberle@zoho.eu, frederic.landragin@ens.fr, todiras@unistra.fr

Abstract

Coreference resolution aims at identifying and grouping all mentions referring to the same entity. In French, most systems run different setups, making their comparison difficult. In this paper, we present an extensive comparison of several coreference resolution systems for French. The systems have been trained on two corpora (ANCOR for spoken language and Democrat for written language) annotated with coreference chains, and augmented with syntactic and semantic information. The models are compared with different configurations (e.g. with and without singletons). In addition, we evaluate mention detection and coreference resolution apart. We present a full-stack model that outperforms other approaches. This model allows us to study the impact of mention detection errors on coreference resolution. Our analysis shows that mention detection can be improved by focusing on boundary identification while advances in the pronoun-noun relation detection can help the coreference task. Another contribution of this work is the first end-to-end neural French coreference resolution model trained on Democrat (written texts), which compares to the state-of-the-art systems for oral French.

Keywords: Coreference, Anaphora, Automatic Coreference Resolution, French, Evaluation

1. Introduction

The coreference resolution task aims at identifying all *mentions* that refer to the same extra-linguistic entity (Pradhan et al., 2012; Recasens and Pradhan, 2016). It may be divided into two subtasks: mention detection and coreference resolution. For example, in Text (1), *Justin Trudeau*, *Canada's prime minister*, *his* and *he* are mentions referring to the same entity (the person named Justin Trudeau). They form a *coreference chain*.

(1) [Justin Trudeau] has won a second term after the country's federal election, but [his] narrow victory means the [Canada's prime minister] will lead a minority government that will be forced to depend on other parties to govern (adapted from *The Guardian*, 11/29/2019).

This task is difficult because of the linguistic and extra-linguistic knowledge required to link various mentions into a single coreference chain (Mitkov, 2002). For example, to find the antecedent of *he*, one must know that *Justin Trudeau* is a person (otherwise the pronoun would have been *it*) and that *Justin* is a male first name. Syntactic knowledge may also prove useful, since both *Justin Trudeau* and *he* are subjects (syntactic parallelism (Poesio et al., 2016; Ng, 2017)). Semantic knowledge, on the other hand, gives the information that *prime minister* is a public title held by a person. Finally, world knowledge may help to find the correct referent, since *Justin Trudeau* is the actual *prime minister of Canada*.

Approaches that address this task automatically may be classified in different ways. For instance, Ng (2017), who focuses on methods used to link mentions to an entity or a coreference chain, differentiates 11 approaches. In this paper, we focus on the *mention-pair*, *easy-first* and *neural network* approaches, because they are particularly important for the French state of the art.

The *mention-pair* approach is well-known: it uses binary classifiers to find pair of mentions that are coreferent, and clusters them later into chain, using transitivity. Example implementations of this approach are found in Ng and Cardie (2002) and Soon et al. (2001). Another approach described by Ng (2017) is the *easy-first*: it builds coreference relations by increasing degrees of difficulty. It is usually implemented by rule-based systems, such as Lee et al. (2013) who propose the use of sieves that detect specific relation types. This approach may require the use of substantial language knowledge, such as syntactic trees, named entity recognition, ontologies, and encyclopedic data (Uryupina, 2007). Finally, the *neural network* approach trains a mention-ranking model that employs a task-specific loss function. Examples include Wiseman et al. (2015), Lee et al. (2017), and Kantor and Globerson (2019). Most of the systems available for French fit in those three approaches.

However, none of these systems have been thoroughly compared to each other, even with the release, in 2013, of a large coreference annotated corpus (Muzerelle et al., 2014), since each system uses slightly different versions of the corpus for evaluation (e.g. different train and test sets). Furthermore, this corpus focuses only on spoken language.

In this paper, we aim to compare performances of different coreference resolution approaches and to evaluate them on both written and spoken French corpora, thus providing an extensive comparison. Additionally, we propose a coreference resolution model that outperforms the current state of the art for French: it is the first French coreference resolution model trained on written texts similar to state-of-the-art systems for English.[1]

This paper is organized as follows: Section 2 presents an overview of French coreference research; Section 3 describes the resources used for our experiments: spoken and written French corpora; and a detailed description of the adaptations we made to the systems, in order to achieve a

[1]The code and models, alongside the enriched corpora are available at `https://github.com/boberle/cofr`.

fair comparison; Section 4 presents our evaluations and our new-state-of the-art-system. Additionally, in Section 5, we discuss the error analysis of our system. The paper closes in Section 6 with some final remarks.

2. Related work

Modern coreference resolution proposals are widely based on large corpora. In English, most of coreference corpora have been developed for evaluation campaigns (Recasens and Pradhan, 2016), such as *Message Understanding Conference* (MUC) (Grishman and Sundheim, 1995; Chinchor, 1998), *Automatic Content Extraction Program* (ACE) (Walker et al., 2006), and CoNLL-2012 (Pradhan et al., 2011; Pradhan et al., 2012). From those, the most used reference corpus for English coreference resolution is CoNLL-2012. It was developed for the CoNLL-2011 and CoNLL-2012 shared tasks, containing 3,493 documents and 1.6m tokens.

Since the 2012 campaign, won by Fernandes et al. (2012) with a CoNLL score of 58.69%, state-of-the-art systems now achieve performances up to 76.6% (Kantor and Globerson, 2019) and 79.6% (Joshi et al., 2019). Both systems are based on (Lee et al., 2017; Lee et al., 2018) (E2E) but use different contextual embeddings: BERT (Devlin et al., 2018) (for which a multilingual version exists) for the first and SpanBERT (Joshi et al., 2019) for the second. Kantor and Globerson (2019) also added an entity equalization mechanism. As originally proposed by Lee et al. (2017), this neural network is an end-to-end coreference resolution model that computes span scores based on span embeddings and a head-finding attention mechanism. In the training step, it maximizes the marginal likelihood of gold antecedent spans from coreference clusters. Later Lee et al. (2018) included a higher-order inference that rules out inconsistent clusters, and a coarse-to-fine model that needed fewer calculations.

In French, there are four corpora dedicated to coreference. The first two (Tutin et al., 2000; Gardent and Manuélian, 2005) are focused on a few linguistic phenomena (e.g. grammatical discourse phenomena for the first, definite descriptions for the second). The other two (ANCOR (Muzerelle et al., 2013; Muzerelle et al., 2014) and Democrat (Landragin, 2016)) are similar in scope to the CoNLL-2012 corpus, as they are aimed at unrestricted coreference and are specifically designed to tackle the coreference resolution task. A remarkable difference between these corpora and the CoNLL-2012 corpus is singleton annotation, present in French corpora and widely used by French systems. In contrast to other mentions, a singleton is a mention that has no coreference relation with any other mention in the document.

Singletons may or not be part of the evaluation setting, and since most of the mentions are singletons (about 80%) they have a significant impact on systems performances (Recasens and Pradhan, 2016), making results including singletons incomparable to results excluding them.

In the remaining of this section, we highlight the most prominent systems for French coreference resolution. Most of them are end-to-end systems.

We start with the most prominent rule-based systems. Trouilleux (2001) uses morphosyntactic information (e.g. grammatical gender and number[2], and some information on the entity type: person, measure, time, etc.). For each pronoun, it searches for possible antecedents, applying a series of rules until one is selected as the correct antecedent. Another rule-based system was proposed by Dupont (2003; Victorri (2005). It is based on salience (Alshawi, 1987) and named entity recognition. It uses syntactic parses and several rules to filter out non-referring expressions, such as dummy pronouns or collocations. More recently, Longo (2013) has developed a system based on Accessibility theory (Ariel, 1990) and morphological and semantic constraints. It begins by detecting entity first occurrences (first mentions in chains) and then build coreference chains by associating other mentions to the first one. ODACR (Oberle, 2019), the last rule-based system presented here, uses syntactic parses and semantic knowledge (e.g. lexical relation and information extracted from Wikipedia). For coreference resolution, it applies linguistic constraints to each pair of mentions. Then, it filters out gender and number-incompatible pairs. Finally, for each remaining pair, ODACR compute a score based on aspects of salience (Lappin and Leass, 1994; Mitkov, 2002), syntactic information (e.g. relative and reflexive pronouns), both syntactic and semantic knowledge (e.g. other pronouns), and semantic knowledge alone (coreference between nouns).

Applications of machine learning approaches for French coreference include Kabadjov and Stepanov (2015), an adaptation of the BART system (Versley et al., 2008), and Godbert and Favre (2017), an hybrid system, that uses CRF (Lafferty et al., 2001) for mention detection and rules for coreference resolution. It uses morphosyntactic and semantic information, such as gender, number and semantic classes (e.g. animacy or vehicle). It detects pronominal anaphora and coreference between noun phrases with identical syntactic head, but not synonyms. CROC (Désoyer et al., 2016) also follows a machine learning approach, but it addresses specifically the mention association task by applying a mention-pair method (Ng, 2017). It uses relational features, that is, features describing the relation between a mention and its candidate antecedents (the distance, whether the mention and the candidate share the same gender, etc.). All these pairs are then classified as coreferent or not. Coreferent pairs are finally associated to form coreference chain. Given that CROC does not detect mentions automatically, its performance is not directly comparable with other systems.

Grobol (2019) has also been trained and evaluated on the ANCOR corpus, like CROC, Kabadjov and Stepanov (2015), and Godbert and Favre (2017). Grobol builds upon the E2E system (Lee et al., 2017) by improving its lack of explicit mention detection and reducing the computational resources required. He proposes the replacement of the mention detection component of the network by another one more suitable to mention detection. He also starts

[2]Grammatical gender and number play an important role in French, since each common noun has an arbitrary gender (feminine or masculine), and one must choose the pronoun according to this gender.

by training only the mention component, and later continues with coreference. Another modification proposed by Grobol is the inclusion of singletons since the original system does not consider them, because they are not annotated in the CoNLL-2012 corpus.

3. Methodology and resources

In this section, we present the (rule-, machine learning-, and deep learning-based) systems and corpora we have selected to compare various approaches to automatic coreference resolution. We used the ANCOR corpus (Muzerelle et al., 2014)), following most of systems for French. But, since it is a corpus of transcribed conversations and interviews, thus focused on speech, and because we aim to provide an extensive evaluation, we also included the DEMOCRAT corpus in our analyses. We detail both corpora in Section 3.1., and then present the systems used in this work in Section 3.2..

3.1. Coreference corpora in French

ANCOR (Muzerelle et al., 2013) is divided into thematic sections: sociolinguistic interviews (complete and incomplete), client-staff dialogues, and phone conversations. Since it is a corpus of spoken language, it has some characteristics that differ significantly from written corpora: there is no punctuation, nor any sentence or paragraph mark, there are speech turns and disfluencies, like word repetitions (e.g. *du du du* "of the the the"), interjections (e.g. *euh* "er"). In order to compare the data, we decided to split all the interviews in the same way, using thematic sections defined by the transcribers of the corpus [3]. The corpus we used thus have 465k tokens, 99k mentions (incl. 50k singletons) distributed across 61k chains, as presented in Table 1.
In 2019, Democrat (Landragin, 2019) was released. It is a large corpus of written French from the 12th to the 21st century. The documents are about 10k word extracts from longer texts (novels, short stories, treatises, biographies, etc.), usually at the beginning. It also contains smaller texts (e.g. press and encyclopedic articles) concatenated into 10k word documents. We kept only the modern part of the corpus (texts from the 19th to the 21st century). We put aside five legal texts, the language of which is very specific, highly repetitive with an unusual syntax (Longo and Todirascu, 2014). We also restored press and encyclopedic articles to their original borders, since they were concatenated in the corpus. The corpus we used, as presented in Table 1, thus have 296k tokens, 82k mentions distributed across 43k chains.
Both Democrat and ANCOR have been divided into *dev*, *train* and *test* sets in the same proportion as the CoNLL corpus (10%, 80%, 10%, respectively).

3.1.1. Annotation scheme differences

Democrat and ANCOR annotation schemes present some differences. This lead us to compute a specific model for each corpus, since the task is not exactly the same (e.g. the system must filter out dummy pronoun for Democrat, but not for ANCOR). This also makes corpus comparison and error analysis difficult since what may be an error in one corpus is just not annotated in the other one (e.g. possessive determiners). The main differences may be summarized as follows:

- Impersonal (or expletive) pronouns (the *il* of *Il pleut* "It is raining") are not considered as referring expressions (Poesio et al., 2016) but they are nonetheless annotated in the ANCOR corpus, as singletons.
- Deictic expressions (as first and second person pronouns, such as *je, tu...* "I, you...", but also address forms like *Monsieur* "Mister") may refer to the same entity. If so, they are coreferenced in Democrat, but not in ANCOR (because the reference is considered to be deictic, i.e. directly to the extralinguistic world). Direct reported speech is frequent in the Democrat corpus, and first and second pronouns are coreferenced to expressions outside the speech, as in: *[Jeanne]$_i$ répondit: "Entre, [papa]$_j$". Et [[son]$_i$ père]$_j$ parut* "[Jeanne]$_i$ replied: 'Enter, [dad]$_j$.' And [[her]$_i$ father]$_j$ appeared."
- Possessive determiners as *mon, ton, son/sa, notre, votre, leur* ("my, your, his/her/it, our, your, their") also refer to the possessor of the word determined, and, as such, are often considered as pronouns (Pradhan et al., 2012; Mitkov, 2002). They are annotated and coreferenced in the Democrat corpus, but not at all in the ANCOR corpus.
- Reflexive pronouns (as *se* in *se laver* "wash oneself") are ignored in the Democrat corpus. In the ANCOR corpus, they are annotated only when they are in the first or second person (singular or plural).
- There are also some differences in mention boundaries. Most notably, the Democrat corpus doesn't include relative clauses inside the limits of the mention.

	Democrat	**ANCOR**
#documents	247	2,124
#tokens	295,978	464,570
#mentions	81,506	98,585
#chains	43,211	60,720
%singleton chains	80.8	83.0

Table 1: Number of document, tokens, mentions, chains and singletons in the Democrat and ANCOR corpora.

3.1.2. Additional annotation

Some of the systems we compare are resource dependent. They use for example named entities and morphosyntactic and syntactic information (e.g. part-of-speech, genre, number, and dependency relation). For a fair comparison, we gave the systems the same set of data for both corpora. This means that we replaced annotated data present in the ANCOR corpus by automatic annotations of the same type as the ones used for the Democrat corpus. For the same reason, we did not used the parsed version of the ANCOR corpus offered by Grobol et al. (2018) (ANCOR-AS), where some interjections (like *euh* "er, uh") have been removed. Both corpora were annotated as follows.
Tokenization, sentence splitting, part of speech, and mor-

[3] This corpus is already transcribed from spoken data, we used it for coreference detection only

phosyntactic data are the result of the *StanfordNLP* tool (Qi et al., 2018). For tokenization and sentence splitting, the process has been semi-automatic: we edited some of the automatic output in order to avoid mentions being split across several sentences, since such mentions cannot be recorded in the CoNLL format used to feed the systems.
Lemma were obtained by looking for surface words in a dictionary (Lefff (Sagot, 2010)). Named entity were annotated by using *Flair* (Akbik et al., 2018), trained on the WikiNer corpus (Nothman et al., 2012). There are 4 types of named entities: person, location, organization and miscellaneous.

3.2. Coreference systems

Our goal is to compare these systems, especially in their coreference task. We have thus split end-to-end systems into two modules: mention detection and coreference resolution. Note that all the systems do not use the same kind of information, as indicated in Table 2. Furthermore, in order to get a fair comparison between these systems and to analyse the errors they produce, we have decided to make them use the same kind of information whenever possible, e.g. same morphosyntactic information and same named entities. We thus have slightly changed the systems described in this section.
We have chosen one system for each of the three coreference resolution approaches:

1. ODACR (Oberle, 2019): a rule-based system based on an easy-first approach;
2. CROC Désoyer et al. (2016): a machine learning-based system that uses the mention-pair model as its coreference resolution approach. This system does not propose a mention identification module, thus we used the mention detection system proposed by Godbert and Favre (2017) (hereafter *GF*), since it is also based on machine learning; and
3. Kantor and Globerson (2019) (hereafter *KG*): a neural network deep learning-based system, which is one of the current state-of-the-art systems for English. We did choose it over the system by Joshi et al. (2019) because BERT has a multilingual version, contrary to SpanBERT.

ODACR originally used parses from Talismane (Urieli, 2013) with the *French TreeBank* part-of-speech tagset, and a built-in, gazetteer-like named entity recognition system with data extracted from Wikipedia (Mahdisoltani et al., 2013). Semantic relations were extracted from a dictionary built from *Glawi* (Hathout and Sajous, 2016). We replaced the parses from Talismane by the dependency parses from *StanfordNLP* (Qi et al., 2018), which uses the *Universal Dependencies* tagset. These changes involved that some level of rules have been adapted. Named entities were identified and tagged with *Flair* (Akbik et al., 2018), trained on the WikiNER corpus (Nothman et al., 2012). We used a French version of WordNet (WOLF) (Sagot and Fišer, 2008; Princeton University, 2010) for semantic relations. Rules have been simplified and weights found for each corpus with a logistic regression (it was originally designed and evaluated for another, ad-hoc corpus). Coreference chains are built by selecting pairs with the highest scores down to a threshold (determined by experiments).
The CROC system, which is concerned only with coreference resolution, applies a SVM model trained over several features available in the ANCOR corpus. The model is trained and evaluated on a balanced subset of mention pairs. However, the coreference prediction phase requires to test all mention combinations. In this scenario, the model lost the expected performance. Instead of SVM, we select the J48 model for CROC, which better fits both scenarios. A second change is related to the features. CROC uses a feature indicating whether a mention is a new entity in the text, i.e. whether a mention is the first in its coreference chain. However, this feature is usually only available when the corpus has been previously annotated or after the coreference resolution task: this is why we removed it. This suppression decreases the CoNLL score by 5% for ANCOR and 1% for Democrat in the development set. To complete the evaluation of the machine learning approach, we needed a mention detection system. We selected GF for this, because it proposed a machine learning-based mention detection.
We trained models for two versions of the system proposed by Kantor and Globerson (2019) (KG). One is the original, strictly end-to-end version, which we have only slightly adapted for our French corpora. The code has been changed to take singletons into account, and to handle document without any mention (this happens with the ANCOR corpus). We also changed the embeddings. Instead of the original large cased Bert embeddings (Devlin et al., 2018) for English (24 layers, 1024 hidden), we used a smaller cased multilingual version (104 languages, 12 layers, 768 hidden). For both word and head embeddings, we used FastText models (Mikolov et al., 2018) for French.
The second version of KG we used is a 2-model split of the system: one for mention detection and one for coreference resolution. For the mentions, we adapted the hyperparameter λ, the proportion of mentions in the corpora ($\lambda = 0.27$ for Democrat and $\lambda = 0.21$ for ANCOR). We also take advantage of the mention loss present in the system, i.e. a loss function guided by the number of correct mentions instead of correct coreference relations. For coreference, we short-circuited the system to introduce pre-detected mentions while keeping all their information in the network, by ordering top mention candidates with our own algorithm.
Democrat is composed by few 10,000 token long texts. In order to be able to run the system, due to resources limitations, we have split long documents into 2k token long chunks. For ANCOR, we split long speech turns (over 100 tokens; some of the original turns were more then 600 tokens in length) at the interjection *euh*. Since there is no sentence boundary annotations in ANCOR, but the system requires the text to be divided into sentences, we used speech turns as sentence boundaries.

4. Results

In this section, we compare the selected systems, both for mention detection (Section 4.1.) and coreference resolution (Section 4.2.). To evaluate coreference resolution without interference from mention detection, systems are fed with gold mentions. In the end (Section 4.3.), we combine

	Tokenizer	Surface	Lemma	PoS	Morphology	Dependency	Paragraph/Speech turn	Speaker	Named entity	Semantic Knowledge	Language Model
Godbert and Favre	✓	✓	✓	✓	✓	✓			✓	✓	
ODACR	✓	✓	✓	✓	✓	✓				✓	
CROC	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Kantor and Globerson	✓	✓						✓			✓

Table 2: Language information and resources required by each system

the best mention detection and coreference resolution systems, aiming to analyse the impact of the full pipeline of the coreference task. Finally, we focus on singletons, evaluating their impact on the best system.

4.1. Mention evaluation

The selected systems obtain significantly different scores. As expected, the KG model outperforms the others, but surprisingly, ODACR and GF present an inconsistent behaviour due to disfluencies in the ANCOR corpus and nested mentions in Democrat. For ANCOR, GF performs better than ODACR, since its rules help to find mention spans.On the contrary, for Democrat, the CRF model has issues for nested mentions.

	Mention Detection (F1)	
	Democrat	**ANCOR**
ODACR	77.35	62.52
GF	67.90	73.32
KG	**88.92**	**88.23**

Table 3: Mention detection evaluation.

4.2. Coreference evaluation

Several metrics have been developed to evaluate automatic coreference resolution systems (see (Poesio et al., 2016) for an overview). We used the official metric of the CoNLL-2012 evaluation campaign, which is the average of three previously published metrics: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAFe (Luo, 2005). We added CEAFm (Luo, 2005) and BLANC (Recasens and Hovy, 2011), which offer a better evaluation of coreference. Concerning these metrics, the most evident conclusion is that KG outperforms the other systems in both corpora (Table 4). Another observation is that systems are ranked according to the different corpora (ODACR $<$ CROC $<$ KG for Democrat and CROC $<$ ODACR $<$ KG for ANCOR). ODACR higher scores with ANCOR probably come from repetitive and less complex chain patterns. Note that MUC evaluates coreference links common to both gold standard and predicted chains. The low scores shown by ODACR and CROC for both corpora, and to a lesser extend KG for ANCOR, indicates a tendency to produce a higher proportion of singletons and shorter chains. On the other hand, a lower BLANC score (which is an evaluation of both coreference and non-coreference relations) with a high MUC score indicates an inability to make good non-coreference decisions, and hence to produce more unwanted relations, as it is the case for KG with Democrat.

4.3. Full-stack end-to-end system

In order to make our results comparable with other end-to-end systems, such as Grobol (2019), we run a full-stack evaluation, by taking the predicted mentions as input to the coreference module. This approach is different from the one originally proposed by Lee et al. (2017), but is similar to the one proposed by Grobol (2019). The result of this evaluation is presented in Table 5 (singleton column). As expected, there is a decrease in the scores; for instance, the CoNLL F1 score decrease from 85.04% to 75.10% in Democrat and from 88.75% to 75.49% in ANCOR. The loss is thus more important for ANCOR; this is because the full-stack evaluation introduces far more mention detection errors than for Democrat (see Section 5.), and these errors have a significant influence on coreference scores.

For Democrat, the performance reduction is about 12% for all metrics except for MUC. On the contrary, for ANCOR, the decrease is about 9% for $CEAF_e$ and $CEAF_m$, about 13% for B^3 and CoNLL, and about 20% for MUC and BLANC. The explanation lies in the absence of effect of singletons on MUC. Indeed, for Democrat, most of the mention detection errors concern singletons (67,7% of errors), hence a lower effect on MUC (which ignores singletons); while for ANCOR fewer errors concern singletons (36,4%).

Another important evaluation parameter is the presence of singletons (82.9% of mentions from ANCOR, and 81.7% of mentions from Democrat), since they may impact the system performances. To address this issue, we evaluate the full-stack system without singletons, i.e. on a corpus without singletons, and after having removed all singletons in the system output. As presented in Table 5, singletons have a strong impact on the results.[4] The impact is stronger on ANCOR (mean of 18%) than on Democrat (mean of 10%). This can be explained by the fact that, while both corpora have a similar proportion of singletons, ANCOR has easy-to-detect singleton mentions: they are mostly first and second person pronouns (*je, tu, nous, vous* "I, you, we"). The average length of singleton mentions is 1.41 tokens. On the other hand, Democrat has longer, 3.34 token-long single-

[4]Note that MUC is not affected by singletons since this measure evaluates only coreference links.

	MUC	B^3	$CEAF_e$	CoNLL	BLANC	$CEAF_m$
			Democrat			
ODACR	30.74	70.07	62.39	54.40	50.66	54.01
CROC	50.52	**84.18**	81.93	72.21	60.28	76.13
KG	**84.13**	83.09	**87.91**	**85.04**	**78.11**	**79.28**
			ANCOR			
ODACR	45.63	83.82	79.97	69.81	60.85	75.11
CROC	37.82	71.13	71.67	60.20	50.39	58.53
KG	**83.08**	**91.44**	**91.74**	**88.75**	**83.97**	**88.49**

Table 4: Coreference score based on the official metrics of the CoNLL-2012.

ton mentions, which are more often nouns with modifiers. Thus, removing singletons disadvantages ANCOR.

	SG		Non SG	
	Democrat	**ANCOR**	**Democrat**	**ANCOR**
MUC	79.03	64.11	79.03	64.11
B^3	71.66	78.40	59.26	57.01
$CEAF_e$	74.35	83.95	59.55	60.36
CoNLL	75.01	75.49	65.95	60.49
BLANC	65.79	63.62	59.07	48.28
$CEAF_m$	69.94	79.18	61.46	63.77

Table 5: Coref score SG vs non SG.

5. Error Analysis

Coreference resolution involves two subtasks: (1) finding referring expressions in the text, (2) clustering them into coreference chains. Coreference errors may originate from each of the subtasks. We will thus analyse mention detection errors and coreference relation errors separately. In this section, we address only the full-stack models presented in Section 4.3. in order to provide a deeper analysis of issues of the model that outperformed the current state of the art. Following the terminology of coreference evaluation metrics (Poesio et al., 2016), in this work, *spurious* mentions or relations are those wrongly predicted by the system (related to precision error) while *missing* mentions or relations are the undetected ones (related to recall error).

5.1. Mention detection errors

Table 6 presents the number of spurious and missing mentions, compared to the total number of mentions in the gold standard corpus (that is, the mentions that should be detected) and the predicted corpus (mentions that the system has actually found). Missing and spurious mentions are roughly in equal proportions, between 10% and 13% of the total number of mentions. This matches the F1 score for mention detection (Table 3).

Mentions	Democrat	ANCOR
Gold	8316	9872
Predicted	8255	10030
Missing	946	407
Spurious	889	1265

Table 6: Missing and spurious mentions in Democrat and ANCOR.

For Democrat, 66.53% of errors involve a noun[5]. This is more than the proportion of nouns found in the gold standard (53.57%), indicating a special difficulty in detecting this kind of mentions. We have identified three problems that may explain noun detection. First, boundaries are harder to detect for longer expressions than for shorter ones like pronouns or possessive determiners, especially at the end since there may be one or several modifiers. Second, noun in lexical collocations or adverbial expressions are not referring, for example expressions of time. Third, the system has difficulties to identify long enumerations of items of a group, frequent in literary texts.

Pronouns are involved in 20.02% of spurious mentions and 8.88% of missing mentions. The proportion is significantly lower than the one in the gold standard (29.69%), so that the system is better at detecting pronouns than nouns, which is not surprising, as pronouns form a limited set of usually single word lexical units. Spurious pronouns are usually impersonal (e.g. *il pleut* "it is raining"), but sometimes result from annotation errors (a pronoun has not been annotated in the reference corpus). Examples of missing pronouns includes compounds units (e.g. *lui-même* "him-self").

For the ANCOR corpus, the most frequent type of erroneous mentions are also nouns, but with a lesser proportion: 46.59% (41.85% in gold standard). Mentions are shorter in ANCOR (mean of 2.46 tokens) than in Democrat (3.24 tokens). Errors are mostly due to speech disfluencies, such as noun repetition at the beginning of a mention.

Missing and spurious pronouns account for 39.83% of all the errors (they represent 52.65% of the mentions in the gold standard). Often, they are demonstrative pronouns (*ce, ça* "that") (30.5%).

5.2. Coreference errors

It is difficult to study coreference errors of an end-to-end system, since it is not possible to fully separate mention misidentifications from coreference issues. However, it allows a better understanding of error source. In this subsection, we attempt to evaluate the proportion of errors due to mention detection and to coreference resolution and study the influence of missing and spurious mentions on coreference errors.

[5]This refers to the part of speech of the syntactic head of the mention. It is not a noun phrase, since neither Democrat nor ANCOR require the annotation of a whole noun phrase (see Section 3.1.)

Here, a relation is a correct, missing or spurious coreference link between two consecutive mentions. We classify incorrect relations into four categories:

1. *coreference only errors* are missing or spurious relations that do not exist in the reference corpus but for which both mentions have been correctly detected;
2. *boundaries errors* are spurious[6] relations due to a boundary detection issue: at least one of the mentions has been detected, but with incorrect boundaries. This requires pairing overlapping mentions. For example, when the system detects *black cat* but the reference mention is *the black cat*, there is an overlap between the mentions, and we pair them together, if the overlapping reference mention is not already paired with another mention;
3. *mixed coreference and boundaries errors* are spurious relations for which we cannot find if they are caused by a boundary or a coreference issue; and
4. *detection errors* are missing or spurious relations for which at least one of the mentions is wrong, because either it is missing or it is spurious and cannot be paired with a reference mention (no overlap).

These error types are interesting because they indicate which part of the system needs to be improved: mention detection (boundary and detection errors) or coreference resolution.

The Table 7 presents error distributions for missing and spurious relations. There are few boundary errors (6.1%), but more for ANCOR than for Democrat, because of repetitions and other speech disfluencies. This is also the reason why detection errors (56% on average), both for missing and spurious relations, exceed coreference only errors. On the other hand, coreference is easier to solve in ANCOR (38.95% of coreference only errors). For Democrat, 23.6% of relation errors are due to detection issues, 73.7% are due to coreference resolution problems.

Relation	Type	Democrat	ANCOR
Missing	*Mention detection*	25.3	59.3
	Coreference	74.7	40.7
Spurious	*Mention detection*	21.9	52.7
	Boundaries	2.5	6.1
	Boundaries or coreference	2.9	4.0
	Coreference	72.7	37.2

Table 7: Missing and spurious relation error types.

5.3. Relation types and distances

Relation types, that is, whether the relation is between a noun and a pronoun, a pronoun and a pronoun, etc., are an important characteristic of coreference, since each relation type has its own properties (see e.g. (Mitkov, 2002; Poesio et al., 2016)). Moreover, the KG system (Kantor and Globerson, 2019) uses few features beside word embeddings: the speaker, the text type[7], and the distance between an antecedent candidate and a mention. Because speaker and text type are not used with Democrat, we have studied missing and spurious relations by their type and distance.

The type is defined by the parts of speech of the two mentions of the relation: *n*oun, *a*jective, adve*r*b, *v*erb, *p*ronoun, possessive *d*eterminer, and other. So, a noun-noun relation would have the type *n-n*. Note that a verb annotation indicates a null subject in Democrat.

For both missing and spurious relation types, we have computed the *criticalness* (i.e. the number of missing or spurious relations of a given type and distance divided by the number of gold or predicted relations) and the *frequency* (i.e. the number of missing or spurious relations of a given type and distance divided by the total number of reference or predicted relations). Correcting criticalness would thus improve quality, while addressing frequency would improve overall performance. For the distance, we have used the 10 semi-log scale buckets defined by Clark and Manning (2016) and used by Kantor and Globerson (2019). This is expressed in number of mentions.

The Figure 1[8] shows spurious relation errors by type and distance, for the Democrat and ANCOR (respectively) corpora (missing relations are not represented here). A more red cell indicates more critical errors for the pair distance and type. For example, a very dark red cell indicates that almost all of the predicted relations are incorrect. This is the case, for example, of noun-determiner relations with no distance between the mentions (Figure 1a). The numbers in the cells are error frequencies. For instance, the noun-determiner with no distance cell shows "1.12": that means that 1.12% of all errors involve noun-determiner relations with no distance between the two mentions.

Consecutive relations with a distance of 0 or 1 mentions are virtually not present (except for some reflexive pronouns in the ANCOR corpus): these kinds of errors are always critical, whatever the type of the relation is. This is however frequent (10.54%) for noun-noun relations in ANCOR; often it is due to the repetition of a word in the same gold mention but divided in the prediction, for example: *les papiers les papiers administratifs*, where *les papiers* ("the papers") is repeated.

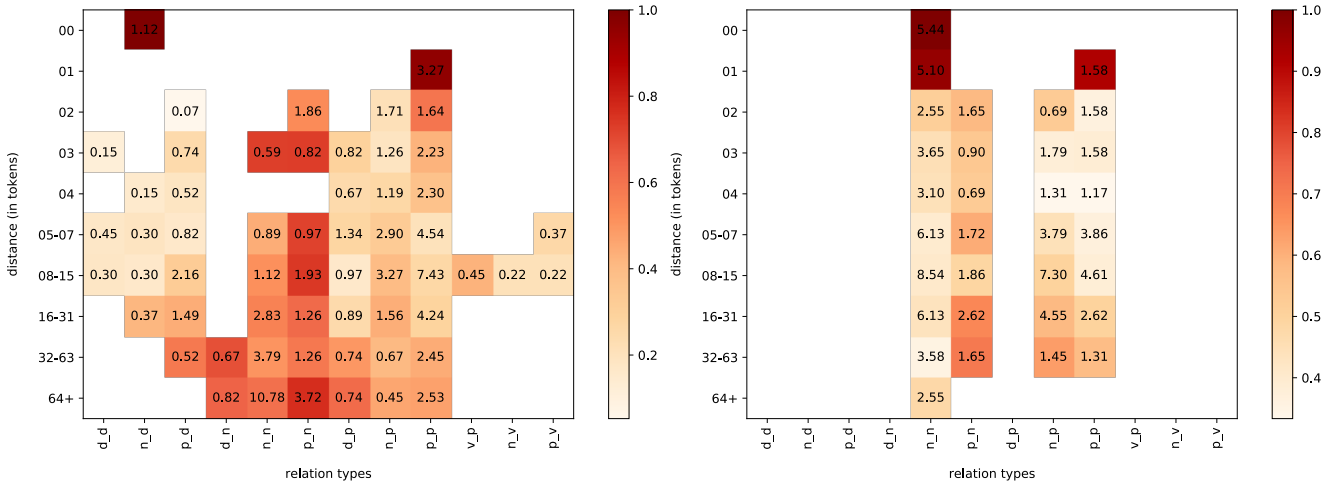
Apart from these values, spurious relation errors in ANCOR are most critical for pronoun-noun, especially at medium distances (0.68 for a distance of 16-31 and 0.71 for a distance of 32-63 mentions). This can be explained by the fact that pronoun-noun relations are less natural than other types since they do not express a kind of antecedent-mention relation because the pronoun could be hardly called an antecedent, and they are never discussed as a valid relation type in the literature (see e.g. (Mitkov, 2002; Corblin, 1995; Poesio et al., 2016). These non-anaphoric relations involving nonetheless an anaphoric element are often found when noun-pronouns relations are isolated in "islands", as in (2):

(2) Cette pauvre mère Chantemesse... elle... elle... ses...

[6]Missing relations involve only gold standard mentions, so this error type affects only spurious relations. This is also the case for the next error type.

[7]For ANCOR only: it is the subcorpus the text comes from

[8]Note that possessive determiners and verbs (null subjects) are not annotated in the ANCOR corpus.



(a) Spurious relation in Democrat (89.7% of the errors are explained). (b) Spurious relation in ANCOR (92.3% of the errors are explained).

Figure 1: Spurious relation error by type and distance. Only relation for which there are more than 10 predictions are shown here.

> elle... elle... elle... *ses*... (81 paragraphs later:) *la mère Chantemesse*... (from Zola, *Le Ventre de Paris*)

As the distance between a pronoun (*ses* in (2)) and the next noun (*la mère Chantemesse*) increases, errors become more critical. This is especially true for Democrat, where texts and chains are longer: pronoun-noun relations with distance of more than 64 mentions are highly critical (.77) and relatively frequent (3.72%). Overall, this column is the most critical.

The most frequent spurious relations are noun-noun: 20% of spurious relations in Democrat, 46.77% for ANCOR. We have seen that this type is problematic for very short distance for ANCOR, because of speech disfluencies. For Democrat, very long distances between mentions are more problematic (10.78%) than short distance relations. This may be due to the fact that Kantor and Globerson's system cut the text into 30-sentence long segments for training: the model is almost never presented with related nouns distant by more than 64 mentions from each other.

For the other types, most of the errors are at medium distances, between 5 and 31 mentions. This is because most of the relations are within this range (54.3% for ANCOR and 45.6% for Democrat).

The missing relation overall picture is similar, with very critical pronoun-noun relations, and most of the errors concerning noun-noun relations. However, there is no issue with very short distances, because they do not exist in the gold corpora, and so cannot be missed.

6. Conclusion

In this work, we compared performances of different coreference approaches for French: mention-pair, easy-first and neural network. The evaluation shows that the neural-based system outperforms the other approaches. In this evaluation, we contrasted written and spoken language. At first, we assessed the mention detection task, and observed no score difference between the corpora. However, the source of the errors differs: for instance, disfluencies (spoken) and mention length (written). Concerning the coreference task, we made two complementary evaluations, one without the influence of mention detection, and one with a full-stack system. The first shows that the neural network performs better in the spoken language corpus, attaching fewer singletons to the wrong chain, well balancing coreferent and non-coreferent mentions. The second shows a strong impact of mention detection errors (with a score decreased by 8.94% for Democrat and 13.26% for ANCOR). In addition, the evaluation without singletons decreases the results by 9.15% (Democrat) and 17% (ANCOR).

Error analysis indicates that most of the errors come from mention detection for ANCOR, but from coreference errors for Democrat. Moreover, pronoun-noun relations are the most problematic, for both corpora.

The next steps include improving the mention detection module and making the system aware of different relation types. Furthermore, we plan to extend the error analysis in order to include more linguistic factors beyond relation type and distance. This will result in a better understanding of which linguistic knowledge is modelled by the network for the coreference resolution task.

7. Acknowledgments

This work was supported by the Alector (*Aide à la lecture pour enfants dyslexiques et faibles lecteurs*) and Democrat projects (*DEscription et MOdélisation des Chaînes de Référence: outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique*), from the French National Research Agency (ANR) (ANR-16-CE28-0005 and ANR-15-CE38-0008, respectively).

8. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alshawi, H. (1987). *Memory and context for language interpretation*. Cambridge Unversity Press.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. Routledge.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.

Chinchor, N. (1998). Overview of muc-7/met-2. In *Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax (1998)*.

Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dupont, M. (2003). *Une approche cognitive du calcul de la référence*. Ph.D. thesis, Université de Caen.

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y., and Dinarelli, M. (2016). Coreference resolution for french oral data: Machine learning experiments with ANCOR. In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2016)*.

Fernandes, E. R., Dos Santos, C. N., and Milidiú, R. L. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 41–48. Association for Computational Linguistics.

Gardent, C. and Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1):115–140.

Godbert, E. and Favre, B. (2017). Détection de coréférences de bout en bout en français. In *Traitement Automatique des Langues Naturelles (TALN'17)*.

Grishman, R. and Sundheim, B. (1995). Design of the MUC-6 evaluation. In *Proceedings of the 6th conference on Message understanding*, pages 1–11. Association for Computational Linguistics.

Grobol, L., Tellier, I., de La Clergerie, E. V., Dinarelli, M., and Landragin, F. (2018). ANCOR-AS: Enriching the ANCOR corpus with syntactic annotations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Grobol, L. (2019). Neural coreference resolution with limited lexical context and explicit mention detection for oral french. In *Computational Models of Reference, Anaphora and Coreference (CRAC)*.

Hathout, N. and Sajous, F. (2016). Wiktionnaire's wikicode GLAWIfied: a workable french machine-readable dictionary. In *International Conference on Language Resources and Evaluation (LREC'16)*.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2019). SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Kabadjov, M. and Stepanov, E. (2015). The sensei discourse analysis tools, 2. techreport, University of Essex.

Kantor, B. and Globerson, A. (2019). Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01: Proceedings of the Eighteenth International Conference on Machine Learning*.

Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, 92:11–15.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Longo, L. and Todirascu, A. (2014). Vers une typologie des chaînes de référence dans des textes administratifs et juridiques. *Langages*, 195(3):79–98.

Longo, L. (2013). *Vers des moteurs de recherche "intelligents": un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence*. Ph.D. thesis, Université de Strasbourg.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2013). Yago3: A knowledge base from multilingual wikipedias. In *7th Biennial Conference on Innovative Data Systems Research*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mitkov, R. (2002). *Anaphora Resolution*. Oxford University Press.

Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E.,

Maurel, D., Villaneau, J., and Eshkol, I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Actes de Traitement Automatique des Langues Naturelles (TALN'2013)*.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). ANCOR centre, a large free spoken french coreference corpus: description of the resource and reliability measures. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'2014)*.

Ng, V. and Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Ng, V. (2017). Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2012). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Oberle, B. (2019). Détection automatique de chaînes de coréférence pour le français écrit: règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques. In *Actes de TALN/RECITAL 2019*.

Poesio, M., Stuckardt, R., and Versley, Y. (2016). *Anaphora resolution*. Springer.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Princeton University. (2010). WordNet.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170. Association for Computational Linguistics.

Recasens, M. and Hovy, E. (2011). BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Recasens, M. and Pradhan, S. (2016). Evaluation campaigns. In *Anaphora resolution*. Springer.

Sagot, B. and Fišer, D. (2008). Building a free french wordnet from multilingual resources. In *Ontolex 2008*.

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Trouilleux, F. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Ph.D. thesis, Université Blaise-Pascal de Clermont-Ferrand.

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S., and Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. In *Proceedings of DAARC 2000*.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Uryupina, O. (2007). *Knowledge acquisition for coreference resolution*. Ph.D. thesis, Universität des Saarlandes.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12. Association for Computational Linguistics.

Victorri, B. (2005). Le calcul de la référence. In Patrice Enjalbert, editor, *Sémantique et traitement automatique des langues*. Hermès.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Wiseman, S. J., Rush, A. M., Shieber, S. M., and Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. 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)*. Association for Computational Linguistics.

9. Language Resource References

Landragin, Frédéric. (2019). *Democrat corpus*. Project ANR Democrat, distributed via Ortolang.

Muzerelle, Judith and Lefeuvre, Anaïs and Schang, Emmanuel and Antoine, Jean-Yves and Pelletier, Aurore and Maurel, Denis and Eshkol, Iris and Villaneau, Jeanne. (2013). *Ancor-Centre corpus*. distributed via Ortolang.