A Dataset for Anaphora Analysis in French Emails

Hani Guenoune ♣ ♢, Kevin Cousot ♣, Mathieu Lafourcade ♢
Melissa Mekaoui ♣, Cédric Lopez ♣
♣ Emvista, Rond-Point Benjamin Franklin, 34000 Montpellier, France
♢ LIRMM, 161 Rue Ada, 34095 Montpellier, France

Abstract

In 2019, about 293 billion emails were sent worldwide every day. They are a valuable source of information and knowledge for professionals. Since the 90's, many studies have been done on emails and have highlighted the need for resources regarding numerous NLP tasks. Due to the lack of available resources for French, very few studies on emails have been conducted. Anaphora resolution in emails is an unexplored area, annotated resources are needed, at least to answer a first question: Does email communication have specifics that must be addressed to tackle the anaphora resolution task? In order to answer this question 1) we build a French emails corpus composed of 100 anonymized professional threads and make it available freely for scientific exploitation. 2) we provide annotations of anaphoric links in the email collection.

1 Introduction

Emails significantly increase the extent of communications in companies. In 2019, about 293 billion emails were sent worldwide every day. They are of great interest for professionals as they represent a valuable source of information and knowledge. Natural Language Processing (NLP) techniques are commonly used for email analysis, to generate a history of the knowledge they convey (Matta et al., 2014), to retrieve redundant problem solving elements (Francois et al., 2016), or to identify tasks in order to help the user manage its time (Khosravi and Wilks, 1999).

Many studies have been conducted on emails since the late 1990s, especially for the analysis of the English language, thanks to the publicly available Enron corpus (250 000 emails sent or received by 87,000 employees of Enron) (Klimt and Yang, 2004).

Our review of French corpora reveals that only one collection of emails is available, which is a subset of the EASY Evaluation Package (provided as part of the Evaluation Campaign for Parsers of French, containing 2,250 anonymised personal emails (Paroubek et al., 2006)). Given this lack of email corpora for the French language, recent works on French emails needed to create their own corpora (Kalitvianski, 2018), However, similarly to the EASY Corpus, the corpora resulting from these works are not freely available to the research community, therefore comparing systems is still unfeasible, as recently highlighted by (Mekaoui et al., 2020).

Anaphora and coreference resolution are core components of the NLP field. These tasks aim to detect and resolve repeated mentions of the same entities in a given document. Many NLP tasks rely on the ability to resolve entities efficiently and could be prominently improved by using robust automatic strategies of anaphora and coreference resolution. In order to design suitable strategies dealing with a natural language phenomenon, one would require to analyse a sufficient number of occurrences of the said phenomenon.

Several studies on annotating French texts with anaphoric links have been conducted (Landragin, 2019; Landragin, 2018; Muzerelle et al., 2014; Tutin et al., 2000), resulting in a few available corpora. The texts covered in these corpora are of various genres and natures, nonetheless, emails are part of none of these datasets.

Proceedings of the 3rd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2020), pages 165–175,
Barcelona, Spain (online), December 12, 2020.

In order to capture the characteristics of referring expressions in this particular kind of discourse (*cf.* Section 3.1), we undertook the process of manually marking the occurrences of each type of anaphora that we considered relevant to the needs of typical NLP issues (*cf.* Section 3.5) in a collection of anonymized professional French emails.

Our work leads to the following contributions: Making available the first free French email corpus for scientific exploitation; Providing annotations for the anaphora discourse phenomenon in French emails; Giving a first quantitative overview of anaphora and coreference in emails.

In this paper, we describe the dataset and its anonymization process (*cf.* section 2), then, we focus on anaphora and coreference annotation (*cf.* section 3). Finally, we describe the annotated corpus and our future works (*cf.* section 5).

2 Dataset

As for (Krieg-Holz et al., 2016), our dataset is made of French professional emails which were requested from individual email authors on the basis of a volunteer act. In our case, authors are employees of a company that wished to remain anonymous. The corpus consists of 100 threads made of 314 emails (7163 words), out of mailing lists, exchanged between June and September 2017.

2.1 Data collection

Technically, the threads were collected through two email inboxes. This implies that the recipients of the threads are always the two same individuals, but emails are from 53 authors. Given the user-centered applications that we considered for this data collection, we made the decision to exclude emails received from automated mailing lists. A thread contains between 1 and 18 emails. Each email consists of the message, the signature if any, and the metadata (”from”,”to”,”subject”, and ”date”). The presence of attachments is indicated. A first cleaning step ensured that no text segments were duplicated in the thread (for instance, through the use of the email forwarding function).

We decided not to structure the threads and share them in their original state, so that the user could have complete freedom over the use of the dataset (no bias according to the structure and no format is imposed). For instance, splitting by sentences or by token is a prepossessing step that embeds several crucial choices we wanted to avoid.

2.2 Anonymization

In order to anonymize the dataset, we used a state-of-the-art named entity recognizer (Lopez et al., 2019) to locate names of people, places, names of organizations, phone numbers, websites, email addresses and so on. Then, all detected mentions have been replaced by dummy data (for instance, ”Peter” could have been replaced by ”Kevin”, and ”Marseille” could have been replaced by ”Paris”). An important point is that the consistence of the corpus is preserved: all identical mentions have been replaced by a given mention. This assures that the anonymizations of coreferent polylexical entites such as ”Marc Sullivan” and ”Marc John Sullivan” remain consistent with the original form of the entities. Moreover, the case has been respected (”Peter” and ”peter” could have been respectively replaced by ”Kevin” and ”kevin”). This allows the corpus to be used for the evaluations of other NLP tasks, such as named entity recognition, for instance. Finally, a manual iteration certifies that no mentions have been left out and that the corpus is fully anonymized. All in all, 9,277 mentions were replaced.

3 Anaphora annotation

Corpora with annotated anaphoric links are essential in NLP and in linguistics. Large sets of annotations give the opportunity to study the anaphora phenomenon. Allowing to craft rule-based systems (manually or automatically generated rules) and machine learning models. Annotated corpora also allow the evaluation of those systems.

The first goal of this study is to address the scarcity of French resources for anaphora study, especially ones containing emails texts (Guenoune et al., 2019). We aim to do so by making available the first free French email corpus with anaphoric links annotations. This work results in a relatively small dataset primarily designed for analysis and evaluation purposes, and represents, in our opinion, a useful starting point (Landragin, 2018). This will allow us to undertake experiments and comparisons to highlight the singularity of automatic anaphora resolution in the email genre (cf. Section 3.1). And thus, serve as a foundation for designing a rule-based resolution system that takes into account the writing conventions observed in this kind of texts.

Two large French corpora are annotated with coreference and anaphoric links. ANCOR (Muzerelle et al., 2014) contains a collection of spoken French transcriptions taken from sociolinguistic interviews. DEMOCRAT's corpus is the most recent resource, annotated with coreferential information (Landragin, 2019). Interestingly, the authors took into account coreference chains in the annotation process. In DEMOCRAT, the selection of texts was done in a way that helps capture the variations of the coreference phenomenon across text genres and eras. Nonetheless, emails are not considered in these corpora.

In Section 4, a set of methodological choices is compared to those of larger-scale projects ; *ANCOR* (Muzerelle et al., 2014), *ARRAU* (Poesio and Artstein, 2008) and *OntoNotes* corpora (Pradhan et al., 2007).

In the following section, we discuss the specificity of annotating anaphoric links in the context of emails (*cf.* section 3.1), then describe our annotation protocol and the typology used (*cf.* section 3.3)).

3.1 Emails singularities

Email writings show some singularities that make the tasks of anaphora annotation and resolution difficult. The two main challenges encountered when dealing with emails are:

- The structural level : Due to the segmented form of the communication (message/thread), emails redefine the binding scope of an anaphoric mention (Reinhart, 1983). An expression can refer to antecedents mentioned within the same message or not. Antecedents may be located in different emails in the thread, or even in different threads. This particular aspect impels us, from an annotation point of view, to design a scheme that handles these extended scopes, and deals with internal and external antecedents. It also affects the resolution task, which becomes analogous to a Cross-Document (CD) problem (Barhom et al., 2019) in which every email thread represents a document. As opposed to Within-Document (WD) anaphora resolution that has been extensively studied during the last decades, the CD task, that aims to locate coreferent entities across multiple documents, remains, as far as we are aware, totally unexplored for French.
  In addition to that, the writings in emails obey to a certain number of stylistic and functional rules making their content singular. One example is the mentions of entities in the metadata of each mail (Sender, recipients, signatures..) that can be antecedents to anaphoric expressions used in the email body.

- The morphological aspects : Similarly to every other kind of user-generated texts, the morphological level affects both the tasks of manual annotation of anaphora and its automatic resolution. For example, gender and number traits being some of the most decisive features in anaphora resolution, rely on meticulous spelling and require a high level of morphological accuracy. However, considering the "non standard" nature of emails writings (Tarrade et al., 2017), morphological errors can produce ambiguous phrasings.

3.2 Annotation protocol

The task was performed by a group of 3 MSc/PhD annotators with Linguistics background, and of different French language proficiency levels. Including one expert annotator whose annotations were considered as gold standard in the agreement study.
The process consisted of six stages (including three iterative steps):

1. **Initialization of the guideline:** The first strategy emerged from discussions about the general purposes and requirements of the resulting corpus.
   The initial draft was defined in such a way that allowed evolution and adaptations to the cases eventually encountered by the annotators.

2. **Selecting a small set of threads:** The goal was to select a portion of the collection that contains the types that are most likely to lead to annotation disparities.

3. **Annotating the selected threads with the current guidelines:** The human annotators were given guidelines regarding the types of mentions and anaphoric expressions to mark. Every annotator trained on a separate portion of the collection.

4. **Agreement study:** In order to assess the operability of the resulting typology, agreement studies have been undertook on the threads selected in step 2. Unlike for other NLP tasks (typically classification ones), annotators are expected to mark words from the text as antecedents, this makes it difficult to establish *a priori* the set of all possible annotations for a given anaphor. Provided that the Kappa measure relies on the set of possible class labels, implementing it to capture agreements on anaphora annotation is challenging. Several methods were used to address this particular aspect (Artstein and Poesio, 2008).
   In our experiments, we chose to isolate the identification and delimitation of antecedents from the classification task. The formal experiments concerned the task of classification of annotated anaphoric mentions into the set of types defined. It was designed to determine in what proportion the annotators agree that an anaphoric markable belongs to a given type, then analyse the reasons for eventual disagreements. Two agreement scores were calculated to assess the consistency of both the typology and the guidelines : A first pairwise (annotator$_i$, gold) agreement score (Cohen's Kappa), and an overall agreement experiment, namely Fleiss' Kappa (Fleiss, J et al., 1981).

5. **Discussion and evolution of the guidelines:** Regular discussions led to several developments in the original typology and annotation protocol (cf. Special operators. Section 3.4). The exchanges mainly dealt with the most commonly encountered difficulties and ambiguous cases. Depending on the pairwise and overall agreement scores, we decide to go back to step 3 or continue to step 6 when a strong agreement is achieved.

6. **Applying the final guidelines to the entire dataset:** We decided to apply the final guidelines to the entire dataset once the Cohen' and Fleiss' Kappa scores (that establish K = 0.6 as good agreement) reached at least 0.70 for the classification task (anaphora types annotation). We chose to keep the typology and the guidelines stable for several sessions, then we decided to proceed with the annotation of the entire collection of threads.

Disagreement

Typically, the cases that resulted in disagreements between annotators concerned the indirect and bridging anaphora phenomena, especially the ones that show through the use of first and second person pronouns in reference to the sender and receiver of the email (this particular case is discussed in Section 5). Multiple disagreements have also been observed in evolving referents annotation (Charolles, 2001) and the inclusive and exclusive use of singular third person pronoun *"On"* (Delaborde and Landragin, 2019). Although not formally assessed, we also note consequent dissension regarding the delimitation of phrasal and abstract antecedents (Amsili et al., 2005).

3.3 Annotation scheme

The major focus in the study was directed towards conceiving a scheme that allows the annotation task to be performed within a reasonable time-frame while maintaining a satisfactory coverage in dealing with the different intricacies of the anaphora phenomenon and consistency with formal specifications.

The anaphoric relation impose a constraint of dependence in interpretation between two distinct mentions, where the first (the antecedent) would be essential to the comprehension of the second (the anaphoric mention) (Mitkov et al., 2012). Therefore, unlike the symmetrical identity (coreference), the relation between an anaphoric mention and its antecedent must be an oriented one. This is represented in the annotation through the choice of a *link-based* strategy. Moreover, as argued in (Poesio et al., 2016), despite the fact that adopting this strategy embeds the necessity to decide which of the antecedents must be marked, it gives the advantage of making the annotation of uncertainty easier.

The strategy used for the annotation task is inspired, in its main aspects, from the guidelines introduced by the MATE Markup Scheme (Poesio, 2005) which seems to us as a straightforward approach to linking lexical units within a text. Slight adjustments have been made in order to maintain a satisfactory level of genericity and to keep the task simple for human annotators. Following the recommendations of the MATE Scheme, we use two distinct elements (`<PHRASE>` and `<ANA>`) to mark discourse entities and anaphoric relations. We used the standard XML format for the final form of the dataset. However, for readability and speed purposes, the annotation was originally performed using a custom tagging language.

This section presents the method used for the identification of different types of mentions, then the representation of the anaphoric relations, with supporting examples in french, followed by their translation.

Entities identification

The annotated markables have been restrained to those involved in anaphoric or coreference relations, for the annotation task to remain manageable and feasible in a reasonable period of time.

The first step is therefore to locate the lexical units that should be linked. The `<PHRASE>` element is used to mark the mentions of discourse entities. It has the numeric attribute `id` that uniquely identifies the mention within the thread. We define the `<PHRASE>` element as the segment that may refer to a concrete or abstract entity of the world, including facts, events or situations. The antecedent syntactical representation could thus be a noun, a verbal phrase or a hole sentence (Amsili et al., 2005). We chose to mark the maximal projection of the head noun. Any modifier, determiner or apposition involved in the description of the entity referred to by the mention, is included in the `<PHRASE>` element, like in the example below.

```
<PHRASE id="1">le nouveau directeur de recherche</PHRASE>

                "The new research supervisor"
```

In addition to nouns' maximal projections, we also mark anaphoric demonstrative and personal pronouns as well as possessive determiners (*cf.* Section 3.5).

Linking referring mentions

Anaphoric relations between mentions are encoded using the `<ANA>` element and are linked to the corresponding `<PHRASE>` elements using the attributes `loc`, `thread`, `src` and `ant` and assigned an anophora type through the attribute `type`, like in the following text, supposedly located in the thread 1.

```
<PHRASE id="1">le nouveau directeur de recherche</PHRASE> s'occupera des
recrutements, <PHRASE id="2">il</PHRASE> fera passer des entretiens dès la
première semaine.
<ANA id="1" loc="I" thread="1" src="2" ant="1" type="PIS"/>

"The new research supervisor will be responsible of recruitment, he will
conduct job interviews beginning the first week"
```

The `loc` attribute takes the values `I`, `E`, or `Et` (respectively for Internal, External and External thread) and indicates whether the antecedent appears in the same message as the anaphoric mention, in a different message of the same thread or in a completely different thread.

The `thread` attribute identifies the thread that contains the antecedent, it allows to retrieve antecedents located in external threads.

Relations are represented by linking the values of the `src` and `ant` attributes of the element `<ANA>` to the `id` of the `<PHRASE>` elements corresponding to the anaphoric mention and its antecedent (resp.).

3.4 Special cases

In order to be able to deal with special forms of anaphoric relations between mentions, a number of special operators have been introduced.

Split antecedents

A split antecedent is an antecedent formed by two separate noun phrases (Chomsky, 1981). A frequent occurrence of this type of anaphoric relation involving a combined antecedent shows through the use of plural pronouns (typically third person ils and elles - ”They”). Since noun phrases involved in a split antecedent can be mentioned in different segments of the text, we chose to identify each one separately with a <PHRASE> element. The combined antecedent is marked in the <ANA> element using the character ”-”.

```
Marc et Eric sont partis, ils étaient pressés. / "Marc and Eric are gone, they were
in a rush."

               <PHRASE id="1">Marc</PHRASE> et <PHRASE id="2">Eric</PHRASE>
               sont partis,<PHRASE id="3">ils</PHRASE> étaient pressés.
               <ANA id="1" loc="I" thread="1" src="3" ant="1-2" type="PIC"/>
```

Dual antecedents

Possessive pronouns [mien, tien, notre...] ”[mine, yours, ours...]” require the identification of two different discourse segments to be fully interpreted. The possessive pronoun relates to a first noun phrase designating the entity that possesses (the owner in the example below). The second dependence is a determining one and binds the pronoun with the phrase that indicates the semantic type of what is possessed, which is omitted from the direct context of the pronoun.

We mark each relation of this phenomenon in the ant attribute of the ANA element, using a separating symbol.

```
               Salut <PHRASE id="1">Rodolphe</PHRASE>,
               <PHRASE id="2">le bureau</PHRASE> de Josette est plus grand que le
               <PHRASE id="3">tien.</PHRASE>
               <ANA id="1" loc="I" thread="1" src="3" ant="1^2" type="PPS"/>

               "Hi Rodolphe, Josette's office is bigger than yours."
```

Chains

Multiple mentions of a given entity are encoded in the element of the anaphoric relation that points to one of the mentions. We try, whenever possible, to mark all previous lexical occurrences of the entity referenced within the ant attribute of the ANA element as follows :

```
     <PHRASE id="1">Pierre Dupont</PHRASE>...<PHRASE id="2">Mr Dupont</PHRASE>...
     <PHRASE id="3">Pierre</PHRASE> ...<PHRASE id="4">Il</PHRASE>.
     <ANA id="1" loc="I" thread="1" src="4" ant="1&2&3" type="PIS"/>

                    "Pierre Dupont...Mr Dupont...Pierre...He.."
```

The strategy of marking only left (previous) coreferents leads to the presence of partial sequences, it means that only the most recent anaphor is linked to a complete coreference chain. This choice has been made in order to avoid the continual correction of preceding annotations manually which would represent an enormous amount of work. In every new apparition, the annotators reuse the previous partial chain and enrich it with the newly mentioned coreferent. This makes the building of chains a sequential procedure that is complete only when the last anaphor of the text has been marked.

Uncertainty

For some special cases, we found it necessary to implement a strategy that deals with fuzzy antecedents. As it happens that the annotator cannot choose unambiguously the correct antecedent between a set of possible mentions. Different methods aim to tackle this issue (Landragin, 2007), we chose to use an annotation with *alternatives*, where the annotator provides a list of all possible mentions separated with ”|”.

This strategy allows us to deal with the case of evolving referents (Charolles, 2001) and the ”non strict” use of singular third person pronoun "On" (Delaborde and Landragin, 2019).

3.5 Anaphoric expressions types

In this section, we begin by listing the syntactic types (parts-of-speech) of the anaphoric expressions we chose to mark. Then, we discuss the semantic types of the relationships between anaphoric mentions and antecedents. Anaphoric expressions retained for annotations are the following.

Pronouns and possessive determiners

Personal, demonstrative (celui, celle... ”the one”) and possessive (*cf.* Dual antecedents, in 3.4) pronouns are selected.

> "As agreed, **they** will offer **you** a new full-time contract with a trial period of one month. The old **one** will be canceled. Charles is in the same case. **He** will receive **his** next week.
> Point to check during the month of September, **your** ability to react responsively to **our** production needs."

Nouns

Nominal anaphora is marked by selecting noun variations that refer to the same entity, in the typical cases, marked nouns are often labels of the semantic class of the named-entity antecedent.

> "I have received a complaint from **FlashDR**, apparently **the client** is not satisfied at all."

Adjectival pointers

A typical case of elliptical phrasing is the use of the attributes of a mention instead of its nominal representation in order to avoid repetition. In the annotation, we target adjectives such as [le premier, le dernier...”the first, the last”] which are used without the noun they are supposed to describe. The determiner is included in the marking element.

> "Here are the two phones in question. For **the first** I wish an estimate, at the same time can you send me a estimate for **the second** ?"

Non-anaphoric forms

This type is provided to annotators, in order to mark phrasings seemingly anaphoric, but should not be considered by a resolution procedure (specifically pronouns in their impersonal use such as the pleonastic *it*, or referential mentions in idiomatic phrases).

3.6 Semantic relations types

The types of anaphoric links we chose to consider are based on the nature of the semantic relation between the two linked mentions.

Identity

The first class of types contains those where the anaphoric mention and the antecedent have a referencial identity, the anaphoric unit points to the whole entity referenced by the antecedent. Coreferent mentions are part of this class.

Association

Often referred to by bridging anaphora, this class contains all non coreference relations between the anaphoric mention and their respective antecedents. Theses semantic relations can be of various natures. It could be a meronymic/holonymic relation where one is part of the entity designated by the second, they could also be linked with a contextual association of ideas. We chose not to distinguish between different semantic relations of bridging anaphora. However, in order not to limit the whole category to the *part-of* relation, we chose to rank every non identity relation within the generic association type.

> "FreeMine has two Bluboo in warranty exclusion in **their** possession, **they** would like to know what **they** should do?"

Synthetic

This type of relation takes place between a phrasal antecedent and the anaphoric mention (mostly demonstrative and personal pronouns or nouns). The pronoun operates a summary of an idea that has been previously described (Lefeuvre, 2012). Anaphora relations involving abstract entities such as ideas, events, fact or situations (Amsili et al., 2005) fall into this category.

```
"Message 1: Our orders are not being validated since yesterday. If it suits you,
we take stock of the problem tomorrow in a meeting.
Message 2: It's ok for me!
Message 3: That's fine with me!"
```

Indefinite pronouns

```
    "If some engineers no longer use our tools, we wonder what they are for!"
```

4 Comparison with other annotated corpora

In this section, we highlight the disparities in methodological choices and conceptual decisions between our annotation protocol and other well-known corpora in the field of anaphora and coreference studies. A synthesis is reported in Table 1 which focus on a series of common issues that are considered important choices to be made in order to design an annotation scheme for anaphora (Poesio et al., 2016).

Corpus	**MD**	**Annotated NPs**	**Predicative NPs**	**Conjunct.**	**Pleo.**	**D-deixis.**
ARRAU	*MaxP+MinA*	*All*	non-referring	*Split*	yes	yes
OntoNotes	*MaxP*	*All*	no	*Coord*	no	no
ANCOR	*MaxP*	*Ana*	no	*Coord*	yes	yes
Ours	*MaxP*	*Ana*	no	*Split*	yes	yes

Table 1: Comparison with ARRAU/OntoNotes

Markables' delimitation (MD). Defining the span of text to be annotated is a consequent question as it raises several issues in the implementation of an evaluation protocol.
In most of the cases, annotation projects chose to mark the maximal projection of the antecedent (noted *MaxP* in table 1). Others, such as the MUC Corpora, add a MIN attribute containing the head of the NP to each markable annotation.
The ARRAU Corpus uses maximal projection and includes the MIN attribute as well (noted *MinA*).

Annotated NPs. In Table 1, we note the distinction between corpora that annotate all NPs and those that choose to mark only NPs involved in anaphoric relations (we use *All* and *Ana* respectively).
For the sake of simplicity and speed of the process, we chose not to mark all NPs.

Predicative NPs. Previous works on annotation diverge on whether predication is to be considered as reference and should or not be linked to the corresponding NP. While several works choose not to not mark them. In ARRAU, predicative NPs are marked but labelled as non-referring.

Conjunction. (Conjunct.) One common issue is to decide whether to mark coordinated NPs (*Coord*) as in "`Mark and Eric`" as NPs. This allows the segment to be linked to a plural third person pronoun ("`They`"). The alternative solution, implemented in this work, is to take into consideration split antecedents (noted *Split* in Table 1), allowing the annotator to link the plural pronoun with two distinct NPs introduced in different segments.

Pleonastic pronouns annotation (Pleo). Whether to consider pleonastic pronouns as markables is another important choice to be made in the annotation process. In our corpus, pleonastic pronouns are annotated as being non referring.

Discourse-deixis annotation (D-deixis). Is a reference to antecedents introduced by phrasal segments, such as references to ideas, events, and abstract objects.

5 Discussion and future work

The analysis of the annotated corpus leads to several axes of study. Referential identity annotation between named entities is complex and raises the question of cross-document coreference. Intuitively, it seems that entities of some semantic types are easier to bind across email threads than others. For instance, it is effortless for a human to decide that two occurrences of the word `"Coca-Cola"`, refer to the same entity, even if each word is mentioned in a separate email thread. On the contrary, it is more challenging to decide whether two mentions of ”`"Bernard"`” appearing in different threads are coreferent or not. A contextual analysis is needed to link such entities.

Another issue deriving from the conversational nature of emails is the question of the identity of the speakers in the metadata, referred to by first person pronouns. The results of our annotation process showed that in a professional setting, plural first and second person pronouns often lead to ambiguous interpretations. For example, an occurrence of the pronoun ”*we*” can refer to the group formed by the sender and the recipients of the email, or to the organisation, mentioned in the signature of the email, to which the speaker belongs.

Using this email corpus we plan to tackle anaphora resolution. The evaluation of the different state-of-the-art machine learning models and symbolic systems will allow us to identify the most significant locks and issues of anaphora resolution in emails.

By taking into account the behaviour of anaphora in such a specific setting, we can focus our efforts towards the challenging cases and the most frequently encountered types of anaphora.

Rel.	***Identity***				***Association***			***Synth.***			***Ellip.***	***Indef.***	***Pleo.***
POS	***PR***	***N***	***Adj***	***Sum***	***PR***	***N***	***Sum***	***PR***	***N***	***Sum***	***Det***	***PR***	***PR***
Count	1008	246	23	**1367**	173	36	**252**	94	20	**114**	**10**	**13**	**103**

Table 2: Typology distribution.

The corpus contains 1856 annotations which gather six relation types (*cf.* Table 2 where possessive determiners have been merged with pronouns under the column *PR* and *N* and *Adj* stand for adjective pointers and nouns respectively). Identity is the most represented relation type (1367 occurrences), followed by association (252) and synthetic (114) relations. As expected, pronominal anaphora is frequent in emails. These links are in most cases internal to the email, as the attached metadata usually defines the antecedent (particularly in the case of first and second person pronouns). In addition to their number, our corpus also contains a high density of pronominal anaphors (74,8%) compared to other corpora ; for instance, the ANCOR corpus contains 41,1% of pronominal anaphors. As a result, only 16,2% of referring nominal mentions are observed, against 45% in ANCOR. Let us note that, interestingly, 90 external relations (identity) have been annotated between two different emails of the same thread, and 23 relations between mentions of separate threads.

6 Conclusion

The detection and resolution of anaphoricity on French emails is an unexplored area that is essential to NLP systems applied to electronic communication. In this paper, we highlight the necessity of building annotated emails corpora. We introduce a small dataset of French emails annotated with anaphoric relations which will be freely distributed. The purpose of the dataset is to make possible the analysis of anaphora's behaviour in an email setting and to serve as a foundation for resolution systems taking into account the writing conventions in this kind of texts. An overview of the distribution of anaphora types in emails gives pointers regarding the challenging aspects of the forthcoming resolution task. We begin by listing the singularities of annotating anaphoric links in emails, then we present the annotation scheme used to deal with special cases and external antecedents. The paper is concluded by comparing important aspects of anaphora annotations to other larger-scale corpora.

References

Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., Dagan, I. (2019). Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution. IN ACL.

Artstein, R., Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. IN Computational Linguistics, 34, 555-596.

Amsili, P., Denis, P., Roussarie, L., and Umr, C. P. (2005). Anaphores abstraites en français : représentation formelle. IN Traitement Automatique des Langues. ATALA.

Charolles, M. (2001). *"Référents évolutifs et évolution de la référence"*. In Les référents évolutifs entre linguistique et philosophie.

Delaborde, M. and Landragin, F. (2019). De la coréférence exacte à la coréférence complexe : une typologie et sa mise en œuvre en corpus. 10èmes Jounées internationales de Linguistique de Corpus, Université Grenoble Alpes, Grenoble, France.

Francois, R., Nada, M., and Hassan, A. (2016). Ktr: an approach that supports knowledge extraction from design interactions. *IFAC-PapersOnLine*, 49(12):473–478.

Guenoune, H., Cédric, L., Tisserant, G., Lafourcade, M., and Mekaoui, M. (2019). Vers une résolution des relations anaphoriques dans la communication électronique médiée. 11.

Kalitvianski, R. (2018). *Traitements formels et sémantiques des échanges et des documents textuels liés à des activités collaboratives*. Ph.D. thesis, Université Grenoble Alpes.

Khosravi, H. and Wilks, Y. (1999). Routing email automatically by purpose not topic. *Natural Language Engineering*, 5(3):237–250.

Klimt, B. and Yang, Y. (2004). Introducing the enron corpus. In *CEAS*.

Krieg-Holz, U., Schuschnig, C., Matthies, F., Redling, B., and Hahn, U. (2016). Code alltag: A german-language email corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2543–2550.

Landragin, F. (2007). L'anaphore à antécédent flou : une caractérisation et ses conséquences sur l'annotation des relations anaphoriques. In Journée d'étude de l'Association pour le Traitement Automatique des LAngues (ATALA) sur la résolution des anaphores. Paris, France.

Landragin, F. (2018). Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus mc4. In Bases, Corpus, Langage - UMR 7320.

Landragin, F. (2020). Rapport final du projet ANR Democrat, "Description et modélisation des chaînes de référence : outils pour l'annotation de corpus et le traitement automatique. ANR (Agence Nationale de la Recherche - France).

Lefeuvre, F. (2012). Les anaphores résomptives en c', cela, ça et ceci dans Juste la fin du Monde et Derniers remords avant l'oubli de Jean-Luc Lagarce, June. Analyse des anaphores résomptives dans des pièces de théâtre de J.-L. Lagarce.

Lopez, C., Mekaoui, M., Aubry, K., Bort, J., and Garnier, P. (2019). Reconnaissance d'entités nommées itérative sur une structure en dépendances syntaxiques avec l'ontologie nerd. In *Extraction et Gestion des Connaissances: Actes de la conférence EGC'2019*, volume 79, pages 81–92. BoD-Books on Demand.

Matta, N., Atifi, H., and Rauscher, F. (2014). Knowledge extraction from professional emails. In *IFIP International Workshop on Artificial Intelligence for Knowledge Management*, pages 43–57. Springer.

Everitt, B. and Fleiss, J. (1981). Statistical Methods for Rates and Proportions.

Massimo Poesio. (2004). The MATE/GNOME scheme for anaphoric annotation.

Chomsky, N. (1981). Lectures on government and binding.

Mekaoui, M., Tisserant, G., Dodard, M., and Lopez, C. (2020). Extraction de tâches dans les emails : une approche fondée sur les rôles sémantiques. In *Proceedings of the Extraction and Gestion des Connaissances (EGC'20)*, page to appear.

Mitkov, R., Evans, R., Orasan, C., Dornescu, I., and Rios, M. (2012). Coreference resolution: To what extent does it help nlp applications? In *TSD*.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol-Taravella, I., and Villaneau, J. (2014). Ancor_centre, a large free spoken french coreference corpus: description of the resource and reliability measures. In *LREC*.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the arrau corpus. In *LREC*.

Poesio, M., Stuckardt, R., and Versley, Y. (2016). *Anaphora Resolution: Algorithms, Resources, and Applications*. ISBN 978-3-662-47908-7.

Pradhan, S. S., Hovy, E. H., Marcus, M., Palmer, M., Ramshaw, L. A., and Weischedel, R. M. (2007). Ontonotes: a unified relational semantic representation. *Int. J. Semantic Computing*, 1:405–419.

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, 6:47–88.

Tarrade, L., Lopez, C., Panckhurst, R., and Antoniadis, G. (2017). Typologies pour l'annotation de textes non standard en français. TALN 2017, June. Poster.

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S., and Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000), 2000, United Kingdom. pp.2

Paroubek,P., Robba, I., Vilnat, A., Ayache, C. (2006). Data, Annotations and Measures in EASY the Evaluation Campaign for Parsers of French. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)", 2006, Genoa, Italy.