# Coreference and Anaphoric Annotations for Spontaneous Speech Corpora In French

Emmanuel Schang[<1>], Aurore Boyer[<2>], Judith Muzerelle[<2>], Jean-Yves Antoine[<3>], Iris Eshkol[<1>], and Denis Maurel[<3>]

[1] Université d'Orléans (LLL)
[2] Université François Rabelais Tours (LLL)
[3] Université François Rabelais Tours (LI)

**Abstract.** This paper presents a corpus-based analysis of coreference and anaphoric relations in French spontaneous conversational speech. It presents the annotation task and two experiments on this corpus (gender and number agreement, definite descriptions as first mention of new discourse entities) which aim at assessing the relevancy of current anaphora solvers on spontaneous speech.

## 1 Introduction

This paper is twofold. It presents: 1) the work done in a pilot study named CO2 which annotates the coreference and anaphoric relations in the ESLO[4] corpus; 2) corpus studies which aim at assessing some features commonly used by NLP anaphora solvers.

In section 2, we present the CO2 project which is a prelude to a larger scale project (named ANCOR[5]) which aims at annotating coreference and anaphora in the ESLO corpus. Section 3 presents the annotation task. In particular, it introduces the features used in the DTD of the annotation tool. Section 4 describes the results of two corpus studies. The first deals with the definite descriptions and the second with gender and number agreements. Section 5 proposes some tracks for complementary studies in this corpus.

## 2 The CO2 project

In the CO2 project presented here, we annotated a corpus of 3 1/2 hours of conversational spontaneous speech. The annotated files are an extract of 35.000 words of the ESLO French corpus, which has been transcribed with Transcriber [2]. Annotation has been done with the GLOZZ annotation tool [10] by one expert and revised by a second one. On the whole, 8910 nominal/pronominal entities and 3513 relations of coreference have been characterized. The annotations are provided as separate XML files which are synchronized on the speech transcripts.

[4] Etude Sociolinguistique d'Orléans: www.univ-orleans.fr\eslo\
[5] The project in its full size (100 hours of annotated dialogs) has been selected for the Région Centre APR-IA 2011's grant and will begin in Autumn 2011.

## 3 The annotation task

### 3.1 Annotating the NPs

The annotation task has two distinct and successive parts:

1. identifying the Named Entities (NE) and broadly, the main elements of an anaphoric chain (pronouns and Noun Phrases).
2. typifying the relations in a chain.

As for the first task, the Named Entities were automatically identified with the finite state transducer cascade CasEN [4] [11]. This tool adopts the ESTER2 conventions [5]. The other NPs (including pronouns) were semi-automatically identified. The tagging of complex NPs was made as follow:

*Example 1.* $[le\,président\,de\,[l'université\,de\,[Tours]\,]\,]$

In this example, each NP can initiate a coreference chain, as in:

*Example 2. Le président de l'université de Tours* est désormais Loïc Vaillant. *Ce dernier* a déclaré qu'il était fier de prendre la responsabilité de *cet établissement.* Cette nouvelle a été chaleureusement accueillie par le maire de *la ville* qui, on le sait, soutenait fortement la candidature du nouvel élu.

Where:

- *Le président de l'université de Tours* is coreferent to *Ce dernier.*
- *l'université de Tours* is coreferent to *cet établissement.*
- *Tours* is coreferent to *la ville*

As for coordinated structures, we choose to identify the group and each member of the structure:

*Example 3.* $[[Pierre]\ et\ [Marie\,Curie]]$

We used GLOZZ.10 [10] to annotate the transcriptions and to annotate the coreferential and anaphoric relations. We customized its DTD to adapt it to our annotation scheme.

### 3.2 The annotation scheme

We followed a detailed annotation scheme, in order to provide useful data to assess the relevance of various linguistic features related to anaphora.

- Part Of Speech:
  - P: Pronouns.
  - N: Nouns (Named Entities NE is a subtype of N)
  - NULL : for artefacts (the NP is split in two between different utterances).

    *Example 4.* U1: Oui alors je voudrais maintenant de la/
    U2: Oui/
    U1: margarine et des œufs
    ' Yes, now I would like some /yes/ margarine and some eggs'

– Grammatical features:
  - Gender (Masc/Fem)
  - Number (Sing/Plur)
– Other features:
  - potential inclusion in a PP.
  - the Named Entities type as defined in the ESTER2 conventions [5].
  - The annotation also describes whether gender and number agreements are found or lacking.

– We retained the criteria for annotating coreference and anaphora presented in [15] for definite and demonstrative NPs, in the wake of Poesio and Vieira's work [13].
  - (Discourse) New: The interpretation of *d* (a definite description) doesn't depend on any previously mentioned expression.
  - Direct Coreference: *d* corefers with a previous nominal expression *a* (its antecedent); *d* and *a* have the same nominal head.

    *Example 5.* La voiture rouge... Cette belle voiture...
    'The red car... This nice car...'

  - Indirect coreference: *d* corefers with a previous nominal expression *a*; *d* and *a* have different nominal heads.

    *Example 6.* Le cabriolet ... cette décapotable ... la voiture...
    'The roadster...this convertible ... the car'

  - Pronominal anaphora (as a special case of the latter).
  - Bridging: *d* does not corefer with a previous expression *a*, but depends for its interpretation on *a*.

    *Example 7.* La voiture... la porte...
    'The car...the door...'

The features used in GLOZZ DTD are summarised in table 1.

## 4 Results

As shown in the previous section, the CO2 corpus is a pilot corpus designed to assess the relevance of several constraints that are usually considered by NLP reference solvers on spontaneous spoken French. The richness of the corpus annotations enables tests that should potentially concern a large range of linguistic features. For the moment, two kinds of experiments have been conducted on the corpus. The first one investigates the nature of the first elements of coreference chains, while the second one focuses on gender and number agreement in coreference relations.

**Table 1.** Customised features of GLOZZ DTD

| Type | |
|---|---|
| Value | Description |
| N | Noun |
| P | Pronoun |
| NULL | Artefact |
| Named Entity | |
| PERS | Humans and pets |
| FONC | Political, military, administrative, etc. functions |
| LOC | Location, place. |
| ORG | Organisations of various types |
| PROD | Human production (films, means of transport, etc.) |
| TIME | Date (duration is in AMOUNT) |
| AMOUNT | Age, duration, weight, etc. |
| EVENT | All kinds of events (Bastille Day, etc.) |
| Definiteness | |
| INDEF | Indefinite |
| DEM | Demonstrative |
| DEF | Definite |
| Relation | |
| Direct Anaphora | See above |
| Indirect Anaphora | See above |
| Bridging | See above |
| Pronominal Anaphora | See above |

### 4.1 Test material

We have conducted several quantitative studies on three annotated files of the CO2 corpus, corresponding to 208 minutes of speech recording and 35192 words. The resulting test corpus includes 8910 nominal or pronominal entities and 3513 relations of co-reference. These anaphoric relations are spread among 550 co-reference chains, which means that a chain includes 6.4 relations on average. In this paper we present some results which are potentially interesting for NLP works in reference solving.

**Table 2.** distribution of the nominal and pronominal entities in the CO2 corpus

| | Discourse New item | Referring item | Total |
|---|---|---|---|
| Nominal entities | 2542 | 1804 (28.9%) | 4346 (49.4%) |
| Pronominal entities | 11 | 4441 (71.1%) | 4452 (50.6%) |

Table 2 shows the distribution of nominal and pronominal entities in the CO2 corpus. Firstly, we notice that nominal and pronominal entities appear quite evenly (49.4% vs. 50.6%). Although a majority of nominal entities introduce a new element of discourse (2542 discourse new entities among 4346 nominal ones), nominal items still represent 28.9% of the co-referring items. This shows that nominal coreferences must be considered by NLP reference solvers, while most works in spoken dialog systems focus only on pronoun anaphora.

**Table 3.** distribution of the references in the anaphoric chains (CO2 corpus)

| | Discourse New references | Other references (inside the chain) |
|---|---|---|
| Nominal entities | 550(99.8%) | 1616 (55.0%) |
| Pronominal entities | 1 (0.2%) | 1323 (45.0%) |

Since there are more pronominal coreferences than nominal ones, it is expected that nominal entities appear more frequently as antecedents in anaphoric chains. Table 3 shows that our observations match partially this conclusion. Even if we only found one unique co-referential chain beginning with a pronoun, pronouns can frequently act as a reference inside an anaphoric chain: they represent 45% of the references in these positions. Consequently, looking uppermost for nouns is not a relevant heuristics for NLP reference solvers in this kind of corpus. Similarly, we have noticed that the antecedent of a referring expression is situated in a prepositional phrase in 27% of the anaphoric relations. Then, searching a reference in a nominal phrase seems relevant but might also be a risky heuristics.

Table 4 presents the distribution of the co-referential relations according to the structural types described in section 3.2. Direct anaphora, which can be easily processed by reference solvers, represents 34.2% of these relations. Pronominal

**Table 4.** Distribution of the anaphora relations in the CO2 corpus

| direct | indirect | pronominal | bridging |
|---|---|---|---|
| 34.2% ($\sigma$ =6.8%) | 15.1% ($\sigma$ =3.5%) | 37.4% ($\sigma$ =4.0%) | 13.4 % ($\sigma$ =7.7%) |

anaphora, which has drown the attention of NLP researchers for years, represent another third of these relations (37.4%).

The resolution of bridging anaphora remains a challenge for reference NLP solvers. Unfortunately, they represent 13.4% of the anaphora attested in the CO2 corpus, which means that their processing cannot be ignored without consequences. Most of these complex coreferences correspond to metonymy.

### 4.2 Definite Description as Discourse New Entities

Definite (as a feature of the noun phrase (NP)) is a feature which is widely considered by coreference solvers (for instance [14]). And it has been already mentioned in various work [13] [6] [9] that definite NPs can introduce new entities in the discourse. The amount of Definite Descriptions (DDs) used as Discourse New entities (DN) in written text has already been studied. [15] presents a ratio of 49.6 % DDs classified as DN in their corpus (the French version of the Official Journal of the EU). A similar amount is found in Portuguese and Brazilian Portuguese texts. In a first experiment, we wanted to evaluate the percentage of DD used as DN in this particular corpus to compare it to the results mentioned in [15] for French. Unsurprisingly[6], the results show that the rate of DD classified as DN in our corpus is strikingly higher: 69,8%.

### 4.3 Gender and Number agreement

Gender and number agreement is a very common constraint which is always considered by any reference solver. It accounts for a mandatory constraint for symbolic solvers (RAP [8]) while being an important feature for heuristic ([12]). While both constraints have proved their usefulness on written language, very few works have tested them on spontaneous speech. Yet, the presence of speech disfluencies and metonymies in conversational speech suggests that one should pay attention to this issue.

We thus have conducted several distributional studies on the CO2 corpus to have a precise picture of gender and number agreement in coreference/anaphora from conversational spoken French. Table 5 presents the results concerning gender agreement.

On the whole, one should consider that gender agreement is well attested in conversational spoken French: 91.3% of the anaphoric relations meet this constraint, as the agreement rate raises up to 99% for direct and pronominal anaphora. This agreement rate decreases significantly (74.5%) with indirect

[6] It is part of the conventional wisdom that Discourse New entities have a different distribution with regard to the genre, see [3] for instance.

**Table 5.** Gender agreement in the anaphoric relations of the CO2 corpus.

| direct | indirect | pronominal | bridging | Total |
|---|---|---|---|---|
| 99.0% ($\sigma$ =1.0%) | 74.5% ($\sigma$ =9.3%) | 98.7% ($\sigma$ =0.8%) | 70.1% ($\sigma$ =9.1%) | 91.3% ($\sigma$ =4.9%) |

coreferences. This was predictable, since gender is quite arbitrary in French : even if two lexical heads describe the same referent, there are great chances that they do not present the same gender. For instance, "voiture" (car) is a feminine word, while its hyperonym "véhicule" (vehicle) is masculine. Gender agreement does not really concern bridging anaphora, since there is no identity of reference between the antecedent and the referring expression in this case (see for instance bridging anaphora with a metonymy). Then, the moderate agreement level (70.1%) that we found is understandable.

In conclusion, this study on the CO2 corpus suggests that conversational spoken French obeys the same constraints as written French as far as gender agreement is concerned: these constraints can usefully be used by reference solvers for direct and pronominal anaphora, but they are not relevant for indirect and bridging anaphora. Our conclusions are slightly different as far as number agreement is concerned. The results presented in Table 6 show that number agreement is significantly less attested than gender agreement in conversational spoken French.

**Table 6.** Number agreement in anaphoric relations of the CO2 corpus.

| direct | indirect | pronominal | bridging | Total |
|---|---|---|---|---|
| 88.3% ($\sigma$ =2.8%) | 85.8% ($\sigma$ =3.9%) | 90.7% ($\sigma$ =5.3%) | 21.9% ($\sigma$ =11.8%)) | 85.3% ($\sigma$ =4.0%) |

On the whole, number agreement is only present in 88.9% of the attested anaphoric relations. This result is consistent with a previous one which only concerned pronominal anaphora [1]. Surprisingly, this moderate agreement holds for every kind of relations. In particular, a noticeable number of direct anaphora do not show number agreement (agreement rate: 88.3%), which was a priori unpredictable. A careful study of the corresponding speech turns shows that in most of these situations, the referent is a generic one. In such case, the plural or the singular can be used indiscriminately in French language, as shown by the following example :

*Example 8.* Sur le plan des honoraires, *les malades* me payent leur consultation et ils sont remboursés à 75%. (...) je n'ai pas le droit de les dépasser, sauf lorsque *le malade* pose des exigences ou s'il s'agit d'une urgence ?

'*the patients* (...) *the patient*'

Such situations may also occur with indirect and pronominal anaphora. For instance, the referring expression le malade in the previous example might have

been replaced without any problem by the indirect anaphoric expression le patient (different lexical head without number agreement) or the singular pronoun *il* (he). This explains the moderate agreement rate we noticed with indirect and pronominal anaphora.

Lastly, number agreement drops down to 21,9% with bridging anaphora. Here, the presence of metonymy is the main explanation for this lack of agreement, as shown by the following example :

*Example 9.* A *l'hotel Caumartin* généralement *ils* sont tous désagréables
'Usually, at *Caumartin Hotel*, *they* are all unpleasant'

We also conducted some additional experiments to assess whether some other linguistic feature might influence number agreement rate. As shown by Table 6, we did not notice a significant influence of any of these features. In all cases (reference in a prepositional phrase, named entity reference, definite or indefinite entity), the agreement rate is situated between 80% and 90%. Number agreement tends to be lower with indefinite reference. This should be explained by the fact that it should correspond more frequently to a generic reference. However, a statistical test shows that the data dispersion is too high (standard deviation $\sigma$ =6.1%) to characterize this decrease as significant. The results in table 7 lead to the same conclusion.

**Table 7.** Number agreement with some specific kind of anaphoric relations.

| Reference in a PP | Named entity reference | Definite reference | Demonstrative reference | Indefinite reference |
|---|---|---|---|---|
| 84.8% ($\sigma$ =6.4%) | 85.2% ($\sigma$ =2.7%) | 87.8% ($\sigma$ =3.8%) | 90.3% ($\sigma$ =7.0%) | 80.4% ($\sigma$ =6.1%) |

To conclude with, this study has clearly shown that number agreement is rather poorly attested in all kinds of co-reference relations. Even though this constraints is met in almost 9 cases out of 10, it would be risky for reference solvers to consider it as mandatory in conversational spoken French. This is why our advice would be to take it into account as a preferential heuristics on spontaneous speech only.

## 5 Future Works

From October 2011, this annotation effort will be continued in a two-year project, ANCOR (Région Centre APR-IA Grant). It will lead to the achievement of an annotated corpus of spontaneous speech including one million words and at least 50 000 coreference relations. It will represent the largest corpus of spoken French with coreference and anaphora annotations. This corpus will be freely distributed and will be of useful for any research on anaphora resolution on

spontaneous speech. In particular, it will enable us to continue the experimental assessment of the linguistic features implemented by anaphora solvers.

Since the recent version of GLOZZ incorporates an inter-annotator agreement tool, we will now be able to calculate the score of agreement on different segments of our corpus and evaluate the strength of the features we used in the experiment.

As we use stand-off annotations (the mark-ups are written in a separated file and they don't overwrite the initial file), the annotations are synchronised on the recordings. This will allow us to work on the physical saliency (see [7]) of the signal and take intonational features in account.

## References


1. Antoine, J.Y.: Résolution des anaphores pronominales : quelques postulats du taln mis à l'épreuve du dialogue oral finalisé. In: Actes TALN2004 (2004)
2. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication 33(1-2), 5–22 (2001)
3. Biber, D., Conrad, S., Reppen, R.: Corpus linguistics: Investigating language structure and use. Cambridge Univ Pr (1998)
4. Friburger, N., Maurel, D.: Finite-state transducer cascade to extract named entities in texts. Theoretical Computer Science 313, 94–104 (2004)
5. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J., Gravier, G.: The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: 9th European Conference on Speech Communication and Technology (2005)
6. Gundel, J., Hedberg, N., Zacharski, R.: Definite descriptions and cognitive status in english: Why accommodation is unnecessary. English Language and Linguistics 5(2), 273–295 (2001)
7. Landragin, F.: Dialogue homme-machine multimodal. Modélisation cognitive de la référence aux objets. Hermès-Lavoisier (2004)
8. Lappin, S., Leas, H.: An algorithm for pronominal anaphora resolution. Computational Linguistics 20(4), 535–561 (1994)
9. Lyons, C.: Definiteness. Cambridge Univ Pr (1999)
10. Mathet, Y., Widlöcher, A.: La plate–forme GLOZZ: environnement d'annotation et d'exploration de corpus. Proc. of 2009 (2009)
11. Maurel, D., Friburger, N., Antoine, J.Y., Eshkol-Taravella, I., Nouvel, D.: Cascades autour de la reconnaissance des entités nommées. TAL 52-1 (2011)
12. Mitkov, R.: Robust pronoun resolution with limited knowledge. In: ACL 98. pp. 869–875. Association for Computational Linguistics (1998)
13. Poesio, M., Vieira, R.: An Empirically Based System for Processing Definite Descriptions. Computational Linguistics 26(4), 525–579 (2000)
14. Soon, W., Ng, H., Lim, D.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27(4), 521–544 (2001)
15. Vieira, R., Salmon-Alt, S., Schang, E.: Multilingual corpora annotation for processing definite descriptions. Advances in Natural Language Processing pp. 721–729 (2002)