Coreference Resolution for French Oral Data: Machine Learning Experiments with ANCOR

Adèle Désoyer, Frédéric Landragin, Isabelle Tellier, Anaïs Lefeuvre, Jean-Yves Antoine, Marco Dinarelli

Coreference Resolution for French Oral Data: Machine Learning Experiments with ANCOR

Adèle Désoyer[1, 2], Frédéric Landragin[1], Isabelle Tellier[1], Anaïs Lefeuvre[3], Jean-Yves Antoine[3], and Marco Dinarelli[1]

(1) Lattice, CNRS, ENS, Université de Paris 3,
Université Sorbonne Paris Cité, PSL Research University, Paris, France
{frederic.landragin,isabelle.tellier,marco.dinarelli}@ens.fr

(2) Modyco, CNRS, Université Paris Ouest – Nanterre La Défense, France
adele.desoyer@gmail.com

(3) Université François-Rabelais Tours, LI, France
{anais.lefeuvre,jean-yves.antoine}@univ-tours.fr

AUTHORS' DRAFT

March 22, 2016

Abstract

We present CROC (Coreference Resolution for Oral Corpus), the first machine learning system for coreference resolution in French. One specific aspect of the system is that it has been trained on data that come exclusively from transcribed speech, namely ANCOR (ANaphora and Coreference in ORal corpus), the first large-scale French corpus with anaphorical relation annotations. In its current state, the CROC system requires pre-annotated mentions. We detail the features used for the learning algorithms, and we present a set of experiments with these features. The scores we obtain are close to those of state-of-the-art systems for written English.

Keywords: mention-pair model, dialogue corpus, coreference resolution, machine learning

1 Introduction

Coreference Resolution has now become a classical task in NLP. This task consists in identifying coreference chains of *mentions* in texts. Supervised machine learning approaches are now largely dominant in this domain, but they require annotated corpora. No such corpus was available for French so far. In this paper, we first describe ANCOR, the first large-scale French corpus annotated with coreferent mentions. It is made of transcribed oral data, which is a specificity relatively to corpora available for other languages. We then present CROC, a baseline system which has been learned with ANCOR, and a variant. Their performances are very close to those observed on coreference resolution challenges for English.

2 The ANCOR Corpus

We present ANCOR, a French corpus annotated with coreference relations which is freely available and large enough to serve the needs of data-driven approaches in NLP. With a total of 488,000 lexical units, ANCOR is among the largest coreference annotated corpora available at present and the only one of comparable size in French.

The main originality of this resource lies in the focus on spoken language. Nowaday systems using NLP for Information retrieval or extraction, for text summarization or even for machine translation have mainly been designed for written language. Oral language presents some interesting specificities, such as the absence of sentence units, the lack of punctuation, the presence of speech disfluencies, and, obviously, the grammatical variability of utterances. See [25] for a more detailed list of oral specific features which make French oral processing a big challenge.

2.1 Presentation of the corpus

The ANCOR corpus is made of spoken French and it aims at representing a certain variety of spoken types. It integrates three different corpora that were already transcribed during previous research projects (Table 1). The first and larger one has been extracted from the ESLO corpus, which collects sociolinguistic interviews [15]. This corpus can be divided in two sub-corpora (ESLO-ANCOR and ESLO-CO2), corresponding to two distinct periods of recordings. It is characterized by a low level of interactivity. On the opposite, OTG and Accueil_UBS concern highly interactive Human-Human dialogues [17]. These two corpora differ by the media of interaction: direct conversation for the first one, phone call for the other one. Conversational speech (OTG and Accueil) only represents 7% of the total because of the scarcity of such free resources in French. All corpora are freely distributed under a Creative Commons license.

Corpus	Speech type	Int.	Size	Duration
ESLO	interview	low	452,000 words	27,5 hours
ESLO_ANCOR			*417,000 words*	*25 hours*
ESLO_CO2			*35,000 words*	*2.5 hours*
OTG	task oriented conversational speech	high	26,000 words	2 hours
Accueil_UBS	phone conversational speech	high	10,000 words	1 hour

Table 1: ANCOR source corpus types and characterization (Int. = Interactivity)

2.2 Annotation scheme

The corpus has been fully annotated by hand on the Glozz platform [14]. Glozz produces a stand-off XML file structured according to a DTD that was specifically designed for ANCOR (it has also been translated in the MMAX2 format for portability purposes). This stand-off annotation allows a multi-layer work and enrichments through time.

The scope of annotation takes into account all noun phrases (NP from now) including pronouns but strictly restricts to them. As a result, the annotation scheme discards coreferences involving verbal or propositional mentions which have been annotated in the OntoNotes corpus. This restriction was mainly intended to favor data reliability by focusing on easily identifiable mentions [11].

Another specificity of the scheme is the annotation of isolated mentions. NPs are annotated even if they are not involved in any anaphoric relation, and this is a real added value for coreference resolution since the detection of singletons is known to be a difficult task [19].

We followed a detailed annotation scheme in order to provide useful data for deep linguistic studies and machine learning. Every nominal group is thus associated with the following features:

- Gender,
- Number,
- Part of Speech (only mentions have been annotated with these features, the corpus does not provide any morpho-syntactic annotation level on its own for other tokens),
- Definition (indefinite, definite, demonstrative or expletive form),
- PP: inclusion or not in a prepositional phrase,
- NE: Named Entity Type, as defined in the Ester2 coding scheme [7],

- NEW: discourse new vs. subsequent mention.

Coders were asked to link subsequent mentions with the first mention of the corresponding entity (discourse new) and to classify the relation among five different types of coreference or anaphora:

- **Direct coreference**: coreferent mentions are NP with the same lexical head.
- **Indirect coreference**: NP coreferent mentions with distinct lexical head (schooner ... vessel).
- **Pronominal anaphora**: the subsequent coreferent mention is a pronoun.
- **Bridging anaphora**: non coreference, but the subsequent mention depends on its antecedent for its referential interpretation (meronomy for instance: the schooner ... its bowsprit).
- **Bridging pronominal anaphora**: the subsequent mention is a pronoun. Its interpretation depends on its antecedent but the two mentions are not coreferent (for instance: the hostel ... they are welcoming).

This annotation scheme is quite similar to previous works on written language [27, 28]. Since ANCOR represents the first large coreference corpus available for French, it is important that the resource should concern researchers that are working on written documents too. Unlike [8], we did not distinguish between several sub-categories of bridging anaphora. We consider such a refined taxonomy to exceed the present needs of NLP while introducing a higher subjectivity in the annotation process. For the same reasons, we did not consider the relation of near-identity proposed in [19]. Recent experiments have shown that near-identity leads to a rather low inter-coders agreement [3].

2.3 Distributional data

This section gives a general outline of the annotated data, to roughly show what should be found in the resource.

Table 2 details how the mentions and relations are distributed among the sub-corpora. With more than 50,000 relations and 100,000 mentions, ANCOR should fulfill the needs of representativity for linguistic studies and machine learning experiments. Table 3 shows that the repartition of nominal and pronominal entities is noticeable stable among the four corpora and leads to a very balanced overall distribution (51.2% vs. 48.8%).

This observation certainly results from a general behavior of French speakers: pronominal anaphora are indeed an easy way for them to avoid systematic repetitions in a coreference chain.

In addition, ANCOR contains around 45,000 annotated Named Entities (NE). Therefore, it should stand for a valuable resource for NE recognition applications. 26,722 NE have been annotated as persons, 3,815 as locations,

Corpus	Number of mentions	Number of relations
ESLO	106,737	48,110
ESLO_ANCOR	*97,939*	*44,597*
ESLO_CO2	*8,798*	*3,513*
OTG	7,462	2,572
Accueil_UBS	1,872	655
Total	116,071	51,337

Table 2: Content of the different sub-corpora

Entities	Nominal	Pronouns	% of NE
ESLO_ANCOR	51.8	48.4	66.3
ESLO_CO2	49.4	50.6	52,4
OTG	47.5	52.5	48.6
Accueil_UBS	48.5	51.5	43.3
Total	51.2	48.8	*59.8*

Table 3: Mentions: distributional information

Person	Location	Organization	Amount	Time	Product
26,722	3,815	1,746	1,496	1,390	1,185

Table 4: Most frequent named entities in ANCOR

Direct	Indirect	Pronominal	Bridging	Bridging pronominal
38,2	6,7	41,1	9,8	1,0

Table 5: Relations: distributional percentages

Corpus	Ancor	CO2	OTG	Accueil_UBS	Total
Direct	41,1	35,2	39,7	40,5	38,2
Indirect	7,3	11,2	6,1	7,5	6,7
Pronoun anaphora	43,9	38,2	46,4	46,0	41,1
Bridging anaphora	10,4	14,4	13,5	11,0	9,8
Pronoun bridging	0,9	1,0	3,3	0,6	1,0

Table 6: Coreference/anaphora: distributional percentages

Agreement	κ	π	α
Delimitation: inter-coder agreement	0.45	0.45	0.45
Delimitation: intra-coder agreement	0.91	0.91	0.91
Type categorization: inter-coder agreement	0.80	0.80	0.80

Table 7: Agreement measures for the ANCOR corpus

1,746 as organizations, 1,496 as amounts, 1,390 for time mentions and 1,185 as products.

Finally, Table 5 presents the distribution of coreference/anaphora relations. Once again, strong regularities between the sub-corpora are observed. In particular, direct coreference and pronominal anaphora are always prevalent. ANCOR contains around 20,000 occurrences of direct coreference and pronominal anaphora which are always prevalent through the corpus.

2.4 Annotation reliability estimation

The estimation of data reliability is still an open issue on coreference annotation. Indeed, the potential discrepancies between coders frequently lead to alignment mismatches that prevent the direct application of standard reliability measures [18, 1, 14]. We propose to overcome this problem by assessing separately the reliability of 1) the delimitation of the relations and 2) the annotation of their types. More precisely, three experiments have been conducted:

1. Firstly, we have asked 10 experts to delimitate the relations on an extract of ANCOR. These coders were previously trained on the annotation guide. We computed, on the basis of every potential pair of mentions, standard agreement measures: κ [4], α [10] and π [21]. This experiment aims above all at evaluating the degree of subjectivity of the task rather than the reliability of the annotated data, since the experts were not the coders of the corpus.

2. On the contrary, the second experiment concerned the annotators and the supervisor of the corpus. We asked them to re-annotate an extract of the corpus. Then we computed intra-coders agreement through a comparison to what they really performed on the actual corpus. This experiment aims at providing an estimation of the coherence of data.

3. Finally, we asked our 10 first experts to attribute one type to a selection of relations that were previously delimited in the ANCOR corpus. We then computed agreement measures on the resulting type annotation.

We observe on table 7 very close results with the three considered reliability metrics (no difference before the 4th decimal). This is not surprising since we consider a binary distance between classes. The inter-coder agreement on delimitation is rather low (0.45). One should however note that this measure

should be biased by our discourse-new coding scheme. Indeed, if a disagreement only concerns the first mention of a coreference chain, all the subsequent relations will unjustifiably penalize the reliability estimation. Further measures to come with the chain coding scheme will soon give an estimation of this potential bias. Anyway, this rather low agreement suggests that the delimitation task is highly prone to subjectivity, even when coders are trained. In particular, a detailed analysis of confusion matrices shows that most discrepancies occur between the delimitation of a bridging anaphora and the decision to not annotate a relation. Besides, this kind of disagreement appears to be related to personal idiosyncrasies. On the contrary, the results become very satisfactory when you consider intra-coders agreement (0.91). This means that our coders followed a very coherent strategy of annotation, under the control of the supervisor. This coherence is, in our opinion, an essential guarantee of reliability. Lastly we observed very good agreements on the categorization task (0.80), which reinforce our decision not to consider near-identity or detailed bridging types.

3 Machine Learning for Coreference Resolution

Coreference Resolution has become a classical task for NLP challenges, e.g. those organized by MUC (`http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html`), ACE (`http://www.itl.nist.gov/iad/mig//tests/ace/`), SemEval (`http://semeval2.fbk.eu/semeval2.php?location=tasks`) or CoNLL (`http://conll.cemantix.org/2011/` & then `http://conll.cemantix.org/2012/`). But none of these challenges included French corpora. For French, as no labelled data were available before ANCOR, only hand-crafted systems have been proposed so far [26, 12]. We rely instead on machine learning approaches. In this section we present our system, named CROC for "Coreference Resolution for Oral Corpus". It only treats the co-reference task. We thus suppose that every mention has already been recognized and associated with its specific features (see section 2.2). The system was trained on the ANCOR_Centre corpus, using the WEKA machine learning platform [29].

3.1 Brief state of the art

Several approaches have been proposed to reformulate coreference resolution as a machine leaning problem. The first and simpler one is the *pairwise* approach which proposes to classify every possible pair of referring mentions as *co-referential* or *not*. This approach assumes that referring mentions are provided (as we do in this paper) and requires a post-processing to build global chains from a set of local pairs. In order to do so, [22, 16] apply a *Closest-First* strategy, which attaches a mention to its closest (on the left) co-referring other mention, whereas [2, 23] propose a *Best-First* strategy, taking into account "co-referential probabilities".

Twin-candidate models [31] are variants of the pairwise approach in which

System	Language	Corpus	MUC	B3	CEAF	BLANC
[22]	English	MUC-7	60.4	—	—	—
[16]	English	MUC-7	63.4	—	—	—
[24]	English	ACE-2003	67.9	65.9	—	—
[23]	English	MUC-7	62.8	79.4	—	—
[9]	English	ACE-2004	67.0	77.0	—	—
[11]	English	CoNNL-2012	68.8	54.56	50.20	—
[12]	French	heterogeneous	36	69.7	55	59.5

Table 8: Results of *end-to-end* systems

System	Language	Corpus	MUC	B3	CEAF	BLANC
[31]	English	MUC-7	60.2	—	—	—
[13]	English	ACE-2	80.7	77.0	73.2	77.2
[6]	English	ACE-2	71.6	72.7	67.0	—
[2]	English	ACE-2004	75.1	80.8	75.0	75.6
CROC	**French**	**ANCOR**	**63.45**	**83.76**	**79.14**	**67.43**
1-Cl. SVM	**French**	**ANCOR**	**61.73**	**84.58**	**80.41**	**69.66**

Table 9: Results of systems starting with pre-annotated mentions

the classification is applied to *triples* instead of pairs: an anaphoric mention and two candidates for its antecedent (the result being either *first* or *second* depending on which of the two candidates is the selected antecedent). criteria between candidates. Other more sophisticated models such as the *Twin-candidate* [31], *mention-ranking* [5] or *entity-mention* [30] have also been proposed. Our coreference resolution system is a baseline, it will thus use the *pairwise* and *Closest-First* strategies.

3.2 Representation of the data in CROC

We have developed CROC as a baseline system which follows the pairwise and closest-first strategies. Pairwise systems rely on a good representation of pairs of mentions. In state of the art models, this representation is usually based on the classical set of *features* proposed in [22], augmented by those of [16]. For our experiments, we used all of these features when they are available in the corpus, plus some new ones we designed. The added features concern speakers and speech turns: they are specific to oral data (in particular to dialogues). One of our purposes is to evaluate the impact of these oral-specific features on the results. For each candidate pair of mentions (i,j), our set of features includes (cf. also table 10):

1. features characterizing each mention i and j:

- at the morphological level: is it a pronoun? is it a definitive SN? is it a demonstrative SN?
- at the enunciative level: is it a new mention?
- at the semantic level: is it a named entity? of which type? Note that no freely available reliable semantic network is available for French, so no other semantic feature was used.

2. *relational* features, characterizing the pair:

   - at the lexical level: are the mentions strictly equal? partly equal?
   - at the morphosyntactic level: do they agree in gender? in number? Note that, in French, even if personal pronouns like "il" (he), "elle" (she)... agree in gender and number with their antecedent, possessive pronouns like "son", "sa"... (his, her...) agree with the noun they introduce and not with the referred antecedent.
   - at the spatial level: how many characters/tokens/mentions/speech turns separate them?
   - at the syntactic level: is one of the mentions included in the other one?
   - at the contextual level: are their preceding/next tokens the same?
   - at the enunciative level: are they produced by the same speaker?

3.3 Baseline results

From the initial corpus, we kept 60% of data for learning, 20% for development, and 20% for test. In order to estimate the influence of the learning corpus size, we distinguished three sets: a small one (71,881 instances), a medium one (101,919 instances) and a big one (142,498 instances). In these sets, 20% of instances are coreferent pairs that are directly extracted from the corpus, and 80% are not-coreferent pairs (negative examples). We also tested different sets of features. In particular, we distinguished three sets: a first one that includes all features, a second one with only relational features, and a third one with all the features that are not linked to oral specificities. A last source of variation concerned the machine learning algorithm used: we tried decision trees, SVM (SMO with default parameters), and Naive Bayes using the Weka platform.

Experiments involving development data showed that the best-performing model is the one calculated by SVM on small training set of data described by all features. Test data are submitted to this model, and the results are filtered by the *Closest-First* method, retaining only the closest antecedent if several pairs involving a mention were found coreferent. We present in tables 8 and 9 the results of some state-of-the-art coreference resolution systems, for the four metrics dedicated to coreference resolution. Oral-specific features do not significantly improve the results.

	Features	Definitions	Possible values
1	m_1_TYPE	syntactic category of m_1	{N, PR, UNK, NULL}
2	m_2_TYPE	syntactic category of m_2	{N, PR, UNK, NULL}
3	m_1_DEF	definition of m_1	{UNK, INDEF, EXPL, DEF_SPLE, DEF_DEM}
4	m_2_DEF	definition of m_2	
5	m_1_GENDER	gender of m_1	{M, F, UNK, NULL}
6	m_2_GENDER	gender of m_2	{M, F, UNK, NULL}
7	m_1_NUMBER	number of m_1	{SG, PL, UNK, NULL}
8	m_2_NUMBER	number of m_2	{SG, PL, UNK, NULL}
9	m_1_NEW	is a new entity introduced by m_1?	{YES, NO, UNK, NULL}
10	m_2_NEW	is a new entity introduced by m_2?	{YES, NO, UNK, NULL}
11	m_1_EN	entity type of m_1	{PERS, FONC, LOC, ORG, PROD, TIME, NO, AMOUNT, UNK, NULL, EVENT}
12	m_2_EN	entity type of m_2	
13	ID_FORM	are m_1 and m_2 forms identical?	{YES, NO, NA}
14	ID_SUBFORM	are there identical sub-forms?	{YES, NO, NA}
15	INCL_RATE	tokens covering ratio	REAL
16	COM_RATE	common tokens ratio	REAL
17	ID_DEF	is there definition equality?	{YES, NO, NA}
18	ID_TYPE	is there type equality?	{YES, NO, NA}
19	ID_EN	is there named entity type equality?	{YES, NO, NA}
20	ID_GENDER	is there gender equality?	{YES, NO, NA}
21	ID_NUMBER	is there number equality?	{YES, NO, NA}
22	DISTANCE_MENTION	distance (number of mentions)	REAL
23	DISTANCE_TURN	distance (number of speech turns)	REAL
24	DISTANCE_WORD	distance (number of words)	REAL
25	DISTANCE_CHAR	distance (number of characters)	REAL
26	EMBEDDED	embedding of m_2 in m_1?	{YES, NO, NA}
27	ID_PREVIOUS	are previous tokens identical?	{YES, NO, NA}
28	ID_NEXT	are next tokens identical?	{YES, NO, NA}
29	ID_SPK	are speakers identical?	{YES, NO, NA}
30	ID_NEW	are discursive status identical?	{YES, NO, NA}

Table 10: CROC complete feature set

3.4 One-class SVM

One of the main problems when using the *pairwise* approach is that, in order to train the binary classification model, artificial negative instances must be generated. Since there is no information to decide whether a pair of not-coreferent mentions is plausible or not, all possible pairs must be generated. The number of such pairs is polynomial in the length of a given mention set in a text, and this in turn means that negative instances are by far more numerous than positive instances. Since this may create a problem of unbalanced representation of positive and negative classes in SVM, heuristics have been proposed to filter out part of negative instances [22, 16]. Despite such heuristics, negative instances are still much more than positive ones.

In order to overcome this problem we investigated the use of models which do not need negative instances. One such model still belongs to the SVM family, namely *One-class SVM* [20]. One-class SVM only needs positive instances, and instead of separating positive and negative instances from each other, separates positive instances from the origin. In order to make a comparison with our baseline, we trained such a model with exactly the same data and features. The results are shown in table 9, line *One-Class SVM*. Our research in this direction is in progress, but we can see that currently results obtained with this approach are roughly equivalent to baseline results.

4 Conclusion and Perspective

Most current researches on coreference resolution concern written language. In this paper, we presented experiments that were conducted on ANCOR, a large French corpus based on speech transcripts, annotated with rich information and coreference chains. This corpus represents the first significant effort to provide sufficient coreference training data in French for machine learning approaches. We described CROC, a baseline approach for automatic coreference resolution on French, as well as another machine learning approach based on one-class SVM. Our first results are roughly equivalent to state-of-the-art performances, which suggests that standard ML approaches for coreference resolution should apply satisfactory on spoken language.

For further investigation, we plan to more carefully study the impact of the various corpus origins on the final results. Does the speech type and/or the level of interactivity influence the way co-reference chains are built in dialogues? To better compare our results with the state of the art, other more complex learning models also need to be tested on these data. And finally, to provide a real *end-to-end* system, we have to automatically identity the mentions and their specific features, as a pre-processing step.

Acknowledgments. This work was supported by grant ANR-15-CE38-0008 ("DEMOCRAT" project) from the French National Research Agency (ANR),

and by APR Centre-Val-de-Loire region ("ANCOR" project).

References

[1] Ron Artstein and Massimo Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, December 2008.

[2] Eric Bengtson and Dan Roth. Understanding the Value of Features for Coreference Resolution. In *Proceedings of EMNLP 2010*, pages 236–243, 2008.

[3] Bartosz Broda, Bartlomiej Niton, Wlodzimierz Gruszczynski, and Maciej Ogrodniczuk. Measuring readability of polish texts: Baseline experiments. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

[4] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[5] Pascal Denis. *New Learning Models for Robust Reference Resolution.* PhD thesis, University of Texas at Austin, 2007.

[6] Pascal Denis and Jason Baldridge. Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 660–669, 2008.

[7] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ester2 evaluation campaign for the rich transcription of french radio broadcasts. In *Proceedings of Interspeech*, 2009.

[8] Claire Gardent and Hélène Manuélian. Création d'un corpus annoté pour le traitement des descriptions définies. *TAL*, 46(1):115–139, 2005.

[9] Aria Haghighi and Dan Klein. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, 2010.

[10] K. Krippendorff. *Content Analysis: an Introduction to its Methodology.* SAGE Publications, Inc., 2004.

[11] Emmanuel Lassalle. *Structured Learning with Latent Trees: A Joint Approach to Coreference Resolution.* PhD thesis, Université Paris Diderot, 2015.

[12] Laurence Longo. *Vers des moteurs de recherche intelligents : un outil de détection automatique de thèmes.* PhD thesis, Université de Strasbourg, 2013.

[13] Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

[14] Yann Mathet and Antoine Widlöcher. Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. In *Actes de TALN*, pages 1–12. ATALA, 2011.

[15] Judith Muzerelle, Anas Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. Ancor_centre, a large free spoken french coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

[16] Vincent Ng and Claire Cardie. Improving Machine Learning Approcahes to Corefrence Resolution. In *Proceedings of ACL'02*, pages 104–111, 2002.

[17] Pascale Nicolas, Sabine Letellier-Zarshenas, Igor Schadle, Jean-Yves Antoine, and Jean Caelen. Towards a large corpus of spoken dialogue in french that will be freely available: the "parole publique" project and its first realisations. In *Proceedings of LREC*, 2002.

[18] Rebecca J. Passonneau. Computing reliability for coreference annotation. In *Proceedings of LREC*, pages 1503–1506, 2004.

[19] Marta Recasens. *Coreference : Theory, Resolution, Annotation and Evaluation.* PhD thesis, University of Barcelona, 2010.

[20] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, 2001.

[21] W. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinions Quaterly*, 19:321–325, 1955.

[22] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[23] Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. Reconcile : A Coreference Resolution Research Platform. Technical report, Cornell University, 2010.

[24] Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 656–664, 2009.

[25] Isabelle Tellier, Iris Eshkol, Samer Taalab, and Jean-Philippe Prost. Pos-tagging for oral texts with crf and category decomposition. *Research in Computing Science*, 46:79–90, Mar 2010.

[26] François Trouilleux. *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. PhD thesis, Université Blaise Pascal, 2001.

[27] Kees van Deemter and Roger Kibble. On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4):629–637, 2000.

[28] Renata Vieira, Susanne Salmon-Alt, and Emmanuel Schang. Multilingual corpora annotation for processing definite descriptions. In *Proceedings of PorTAL*, 2002.

[29] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical machine learning tools and techniques with java implementations, 1999.

[30] Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL'08*, pages 843–851, 2008.

[31] Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. Coreference resolution using competition learning approach. In *Proceedings of ACL'03*, pages 176–183, 2003.