ANCOR_Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures

Judith Muzerelle[1], Anaïs Lefeuvre[2], Emmanuel Schang[1], Jean-Yves Antoine[2], Aurore Pelletier[1], Denis Maurel[2], Iris Eshkol[1], Jeanne Villaneau[3]

[1] Univ. Orléans, LLL UMR 7270
6 Avenue du Parc Floral, 45100 Orléans
[2] Université François Rabelais de Tours, LI
64 Avenue Jean Portalis, 37200 Tours
[3] Université Européenne de Bretagne, IRISA
Rue Yves Mainguy, BP 573 56017 Vannes cedex

E-mails: judith.muzerelle@etu.univ-tours.fr, anais.lefeuvre@univ-tours.fr, emmanuel.schang@univ-orleans.fr, jean-yves.antoine@univ-tours.fr, aurore.pelletier-2@etu.univ-tours.fr, denis.maurel@univ-tours.fr, iris.eshkol@univ-orleans.fr ,jeanne.villaneau@univ-ubs.fr

Abstract

This article presents ANCOR_Centre, a French coreference corpus, available under the Creative Commons Licence. With a size of around 500,000 words, the corpus is large enough to serve the needs of data-driven approaches in NLP and represents one of the largest coreference resources currently available. The corpus focuses exclusively on spoken language, it aims at representing a certain variety of spoken genders. ANCOR_Centre includes anaphora as well as coreference relations which involve nominal and pronominal mentions. The paper describes into details the annotation scheme and the reliability measures computed on the resource.

Keywords: coreference, French spoken language, free annotated corpus

1. Introduction

Information Retrieval and documents indexing should certainly be accounted for one of the most promising application areas of Natural Language Processing (NLP). These tasks require a large resort to NLP treatments, among which the resolution of coreference and anaphoric relations plays a crucial role. It allows indeed various lexical mentions that are present in texts to be clustered with the single discourse entity they refer to. Several evaluation campaigns (MUC, ACE, SemEval) have demonstrated the existence of operative solutions to coreference resolution, which is nowadays extended to the question of entity linking (Rao et al., 2011). However, the lack of large annotated corpora still restricts the achievement of efficient resolution systems. In this paper, we present ANCOR_Centre (ANCOR as an abbreviated form), the first French corpus which is freely available and is 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 corpora available at present[1] (Table 1).

Language	Corpus	Gender	Size (words nb.)
German	TüBa-D/Z *(Hinrichs et al., 2005)*	News	800,000
English	OntoNotes *(Pradhan et al., 2007)*	Various genres: news, conversational phone calls weblogs, use net, broadcast, talk shows	500,000
Chinese	OntoNotes *(Pradhan et al., 2007)*		400,000
Catalan	AnCora-*Ca (Recasens & Marti, 2010)*	News	400,000
Spanish	Ancora-Es (*Recasens, 2010*)	News	400,000
Japanese	NAIST Text *(Idia et al., 2007)*	News	970,000
Dutch	COREA *(Heindrickx et al., 2008)*	News, spoken language, encyclopedias	325,000
Czech	PDT *(Nedouluzhko et al., 2009)*	Newspaper	800,000
Polish	PCC (*Ogrodniczuk et al., 2014*)	various written and spoken genders	514,000

Table1 – Largest manually annotated coreference corpora (more than 325,000 words)

[1] The DEDE corpus (Gardent & Manuelian, 2005), which was until now the largest coreference corpus freely available on French, focuses only on definite description. It sizes restricts to 48,000 words.

To the best of our knowledge, ANCOR represents the largest corpus that concerns specifically spoken language. ANCOR follows a rich annotation scheme to fulfil the needs of NLP machine learning as well as linguistic investigations.

This paper presents into details this resource. First, we present the speech corpora on which the annotation was conducted. We then describe our annotation procedure. Section 3 gives some distributional information on the data that are present in the corpus. The next section addresses the question of the estimation of data reliability on coreference annotation and gives the results obtained on the ANCOR corpus. In conclusion, we discuss the availability of the corpus.

2. Source spoken corpora

The ANCOR corpus focuses exclusively on spoken French. Although it cannot be considered as a balanced corpus, it aims at representing a certain variety of spoken types. It consists of four different spoken corpora that were already transcribed during previous research projects (Table 2). Two of them have been extracted from the ESLO corpus, which collects sociolinguistic interviews with a restricted interactivity (Schang et al., 2012). On the opposite, OTG and Accueil_UBS concern highly interactive Human-Human dialogues (Nicolas et al., 2002). These last two corpora differ by the media of interaction: direct conversation or phone call. All of these corpora are freely distributed under a Creative Commons license. Conversational speech only represents 7% of the total corpus because of the scarcity of such free resources in French.

Corpus Parole	**Speech type**	**Words number**	**Duration**
ESLO_ ANCOR	Interview	417,000	25 hours
ESLO_ CO2	Interview	35,000	2.5 hours
OTG	Task-oriented conversational speech	26,000	2 hours
Accueil_UBS	Phone conversational speech	10,000	1 hour

Table 2 – Source corpora of ANCOR_Centre

3. Annotation procedure and annotation scheme

Although we have conducted some experiments on the automatic detection of nominal groups and named entities on the ESLO corpus, we finally decided to fully annotate by hand these corpora on the GLOZZ platform (Mathet and Widlöcher, 2012). Glozz produces a stand-off XML file structured after a DTD that was specifically designed for ANCOR. This stand-off annotation allows a multi-layer work on the data and potential enrichments through time.

In order to restrict the cognitive load of the coders and to favour intra-coder coherence, the annotation process was split into four successive phases:

1. Mention borders marking (coders: Master or PhD students in linguistics)
2. Adjudication of phase 1 by a super-annotator
3. Marking of coreference or anaphora relations (same coders)
4. Adjudication of phase 3 by a super-annotator

The scope of annotation covers all noun phrases including pronouns but restricts strictly to them. For instance, a noun phrase like *le lendemain* (*the day after*) is considered a legitimate mention, while the adverbial *demain* (*tomorrow*) will be ignored. This precise delimitation favours the annotation reliability since it provides coders with objective rules to characterize what should be considered or not during the annotation. As a result, the annotation scheme discards coreferences involving verbal or propositional mentions. These relations contain abstract anaphoras, which are beyond the aims of our project and would have required a very specific annotation scheme (Dipper and Zinmeister, 2010).

We follow 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,
- 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 (Galliano et al., 2009),
- NEW: discourse new mention vs. subsequent mention.

There is no real consensus on the way coreferent mentions should be related in the annotation. In the ANCOR project, we asked coders to link always subsequent mentions with the first mention of the corresponding entity (discourse new). Alternative coding schemes have however their own relevancy. This is why the corpus is distributed with three alternative representations:

- **Discourse-new coding scheme**: relations from subsequent to first mentions,
- **Coreference chain coding scheme**: relations from one coreferent mention to the next one,
- **Cluster coding scheme**: sets of coreferent mentions.

Marked relations are additionally classified among five different types of coreference or anaphora:

- **Direct coreference**: coreferent mentions are NP with the same lexical head.
- **Indirect coreference**: coreferent mentions are NP 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**: specific bridging anaphora where the subsequent mention is a pronoun. We distinguished this type in order to emphasize metonymic situations (*Avoid the Grand Central Hotel ... they are unpleasant*) which occur frequently in conversational speech.

This annotation scheme is quite similar to previous works on written language (van Deemter & Kibble, 2000, Vieira et al., 2002). 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 (Gardent and Manuélian, 2005), we didn't 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 didn't consider the relation of near-identity proposed in (Recasens, 2010). Recent experiments have shown that near-identity leads to a rather low inter-coders agreement (Ogrodniczuk et al., 2014). Section 5 details the data reliability measures obtained on our corpus.

4. Corpus description: distributional data

Although ANCOR clusters valuable information for deep linguistic analyses, this section gives only a general outline of the annotated data, to show roughly what should be found in the resource[2].

Corpus	Nb. of mentions	Nb. of relations	*Mention/ relation ratio*
ESLO_ANCOR	97,939	44,597	*2.19*
ESLO_CO2	8,798	3,513	*2.50*
OTG	7,462	2,572	*2.90*
Accueil_UBS	1,872	655	*2.86*
TOTAL	116,071	51,337	*2.26*

Table 3 – Content of the different annotated sub-corpora

Table 3 details the distribution of the mentions and relations among the sub-corpora. With more than 50,000 relations and 100,000 mentions, ANCOR should fulfil the needs of representative linguistic studies and machine learning.

Table 4 shows that the repartition of nominal and pronominal entities presents a noticeable stability among the four corpora and leads to a very balanced overall distribution (51.2% vs. 48.8%).

[2] People interested in a qualitative approach of the resource can consider (Lefeuvre et al., 2014) for an illustration of comprehensive linguistic studies that should be conducted on the resource.

Corpus	Nominal entities	Pronouns	*% of Named Entities*
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 4 – Mentions: distributional information

This observation results certainly from a general behaviour of spoken French: pronominal anaphora is indeed an easy way for French speakers to avoid systematic repetitions in a coreference chain. On the contrary, the use of Named Entities (NE)[3] is strongly related to the discourse domain. This explains that we observe significant variations of their relative frequency from one corpus to another[4]. ANCOR clusters around 45000 annotated Named Entities (table 5) Therefore, it should stand for a valuable resource for named entities recognition applications.

PERS	LOC	ORG	AMOUNT	TIME	PROD
26,722	3,815	1,746	1,496	1,390	1,185

Table 5 – Most frequent named entities in ANCOR (number of occurrences – Ester2 Types)

Corpus	ESLO _ Ancor	ESLO _ 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 information

Finally, Table 6 presents the distribution of coreference/anaphora relations. Once again, strong regularities between the sub-corpora are observed. In

[3] Following the ESTER 2 Evaluation Campaign (Galliano et al. 2009), we have used 7 categories: PERS stands for person, LOC for location, ORG for organization, AMOUNT for amount, TIME for time, PROD for product, in addition to FUNCT for functions which represents 0,5 % of the NE in the corpus.

[4] For instance, locations (LOC) represent only 8,1% of the observed named entities in the Accueil_UBS corpus, while this ratio increases up to 19,4% on the OTG one: tourists information involves indeed frequent references to locations.

particular, direct coreference and pronominal anaphora are always prevalent. ANCOR clusters around 20,000 occurrences of these two relations.

5. Annotation reliability estimation

The estimation of data reliability is still an open issue on coreference annotation. Indeed, the potential discrepancies between coders lead frequently to alignment mismatches that prevent the direct application of standard reliability measures (Passoneau, 2004; Artstein & Poesio, 2008 ; Matthet & Widlöcher, 2011). 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've 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 all potential pair of mentions, standard agreement measures: κ (Cohen, 1960), α (Krippendorff, 2004) and π (Scott 1955). 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.

Corpus	Kappa	Pi	Alpha
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

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 (Antoine et al., 2014). 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 concerns only 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 give soon 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 categorisation task (0.80), which reinforce our decision not to consider near-identity or detailed bridging types.

6. Conclusion: corpus avalaibility

The inter-coder agreement observed on the ANCOR corpus suggests that it represents a reliable annotated resource. With a size approaching 500,000 words, ANCOR has no equivalent for French and represents one of the largest coreference corpora on spontaneous speech. It is freely distributed under a CC-BY-NC-SA Creative Commons licence. The version 1.0 of the resource, which only handles a discourse-new coding scheme, can be downloaded from the webpage of the ANCOR project (http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html). The final version of ANCOR (three coding scheme) will be available on the *Speech and Language Data Repository* (http://crdo.up.univ-aix.fr) on mid-2014.

7. Acknowledgments

The development of the ANCOR_Centre corpus was conducted in the framework of the CO2 project (joint founding of Orléans U. and François Rabelais Tours U.) and mainly of the ANCOR project, founded by the Centre Region.

8. References

Antoine, J.-Y., Villaneau, J., Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi- coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. *Proc. EACL'2014*. (to appear)

Artstein, R., Poesio, M. (2008) Inter-Coder agreement for Computational Linguistics, *Computational Linguistics*, 34, pp .555--596.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pp. 37--46.

Dipper S., Zinmeister H. (2010). Towards a standard for annotating abstract anaphora. *Proc. LREC'2010 workshop on Language Resources and Language Technology Standards*. Valetta, Malta, pp. 54--59.

Galliano, S., Gravier, G., Chaubard, L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcast. *Proc. Interspeech 2009*, Brighton, UK.

Gardent, C. and Manuelian, H. (2005). Création d'un corpus annoté de traitement des descriptions définies. *Traitement Automatique des Langues, TAL* , 46(1).

Heindrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Van Der Vloet, J., Verschelde, J.-L. (2008). A coreference corpus and resolution system for Dutch. *Proc. LREC'2008*.

Hinrichs, E., Kübler, S., Naumann, K., Zinsmeister, H. (2005). Recent developments in linguistic annotations of the TüBa-D/Z Treebank. *Proc. 27th Annual Meeting of the German Linguistic Association*, Köln, Germany.

Iida, R., Mamoru, K., Kentaro, I., Yuji, M. (2007). Annotating a Japanese text corpus with predicate-argument and coreference relations. *Proc. Linguistic Annotation Workshop (LAW 2007)*,Stroudsburg, PA, USA. ACL, pp. 132--139.

Krippendorff, K. (2004). *Content Analysis: an Introduction to its Methodology*. Chapter 11. Sage: Thousand Oaks, CA.

Passoneau, R. (2004). Computing reliability for Co-Reference Annotation. *Proc. LREC'2004*. Lisboa, Portugal.

Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., Micciula, L. (2007). Unrestricted coreference: identifying entities and events in OntoNotes. *Proc. 1st IEEE Int. Conf. on Semantic Computing (ICSC'07)*. Washington, DC. USA. IEEE, pp. 446--453.

Lefeuvre, A., Antoine, J.-Y., Schang, E. (2014). Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé. *Proc. CMLF'2014*. (to appear)

Mathet, Y., Widlöcher, A. (2012). The Glozz Platform: A Corpus Annotation and Mining Tool. *Proc. 2012 ACM symposium on Document engineering*, pp. 171--180.

Nedoluzzhko, A., Mírovský, J., Ocelák, R., Pergler, J. (2009). Extended coreference relations and bridging anaphora in the Prague Dependancy Treebank. *Proc. DAARC'2009*. Chennai Goa, Indica , pp. 1--16

Nicolas, P., Letellier-Zarshenas, S., Schadle, I., Antoine, J.-Y., Caelen, J. (2002). Towards a large corpus of spoken dialogue in French that will be freely available: the "Parole Publique" project and its first realisations. *Proc. LREC'2002*. Las Palmas de Gran Canaria, Spain.

Rao, D., MacNamee, P., Dredze, M. (2011). Entity Linking: Finding Extracted Entities in a Knoledge Base. In T. Poibeau, H. Saggion, J. Piskorski, R. Yangarber (Eds.), *Multi-source, Multilingual Information Extraction and Summarization*. Springer, pp. 93--115.

Recasens Potau, M., (2010). Coreference: Theory, Annotation, Resolution and Evaluation. PhD Thesis, Universitat de Barcelona, Catalunya, Spain.

Schang, E., Boyer, A., Muzerelle, J., Antoine, J.-Y., Eskhol, I., Maurel, D. (2011). Coreference and anaphoric annotations for spontaneous speech corpus in French. *Proc. DAARC'2011*, *Discourse Anaphora and Anaphor Resolution Colloquium*, Faro, Portugal.

Ogrodniczuk, M., Kopeć M., Głowinska K., Savary A., Zawisławska, M. (2013). Polish coreference corpus, submitted at *LTC'2014*.

Scott, W. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinions Quaterly*. 19, pp. 321--325.

van Deemter, K., Kibble, R. (2000). On Coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics,* 26(4), pp. 629--637.

Vieira, R., Salmon-Alt, S., Schang, E. (2002). Multilingual corpora annotation for processing definite descriptions. *Proc. Advances in Natural Language Processing 2002*, pp. 721--729.

.