ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations

Loïc Grobol*†, Isabelle Tellier*, Éric de la Clergerie†, Marco Dinarelli*, Frédéric Landragin*
* Lattice ; † ALMAnaCH, Inria
1 rue Maurice Arnoux, 92120 Montrouge, France ; 2 rue Simone Iff, 75589 Paris, France
{loic.grobol, isabelle.tellier, marco.dinarelli, frederic.landragin}@ens.fr,
eric.de_la_clergerie@inria.fr

Abstract

This paper presents ANCOR-AS, an enriched version of the ANCOR corpus. This version adds syntactic annotations in addition to the existing coreference and speech transcription ones. This corpus is also released in a new TEI-compliant XML format.

Keywords: Coreference, Parsing, Oral Data, TEI

1. Introduction

Since its inception, even as far back as (Hobbs 1986), automatic detection of coreference and anaphora has made consequent use of rich syntactic knowledge. To this day, most of the state-of-the-art systems make use of at least some information inferred from syntax analyses available in coreference-annotated treebanks. Until now, however, no French coreference corpus had publicly available syntactic analysis, which led previous works on automatic coreference detection to either use ad-hoc automatic parsing and shallow syntactic analysis or to use manually annotated mentions, thus avoiding the problem of mention detection.

In an effort to at least start to address this lack of resource, we present ANCOR-AS, an enriched version of ANCOR (Muzerelle et al. 2014) — the current reference corpus for coreference in French — that includes syntactic analysis obtained in the Universal Dependencies framework (Nivre et al. 2016) through state-of-the-art automatic parsing techniques. ANCOR-AS builds upon our previous efforts in (Grobol, Landragin, and Heiden 2017) to develop a standard format for coreference annotations and extends our proposal for TEI-compliant (TEI consortium 2016) reference annotations to the case of syntactic dependencies.

Our preliminary experiments with using automatic syntactic analysis for mention detection, described in (Grobol, Tellier, et al. 2017) give us confidence that in the absence of gold syntactic analysis, automatic parsing is a valuable asset for developing an end-to-end coreference detection system. Moreover, unrelated works on ANCOR, such as Temporal@ODIL (Antoine et al. 2017), are already using semi-automatic parsing as a support for further manual annotations, which suggests that standardized syntactic annotations would be useful for further uses of ANCOR beyond coreference.

2. Context

2.1. Related works

As mentioned earlier, there already exist several coreference treebanks, most notably the AnCoRa corpus (Taulé, Martí, and Recasens 2008) (for Spanish and Catalan), the Prague Dependency Treebank (Nedoluzhko et al. 2016) (for Czech) and the omnipresent OntoNotes (Pradhan, Hovy, et al. 2007). To this day, the latter is still the largest coreference corpus, with nearly 3M words in three languages – Arabic, English and Chinese – and the evaluation standard for coreference detection system since its use for the CoNLL-2011 (Pradhan, Ramshaw, et al. 2011) and CoNLL-2012 (Pradhan, Moschitti, et al. 2012) coreference detection shared tasks. Also relevant for our case is the NXT-Switchboard corpus (Calhoun et al. 2010), a coreference treebank of oral English transcriptions.

For some other languages, only coreference corpora with rich morphological or shallow syntactic annotations exist, such as the Polish Coreference Corpus PCC (Ogrodniczuk et al. 2015) or the EPEC corpus (Soraluze et al. 2012) (for Basque) and, more recently, the Summ-it++ corpus (Fonseca et al. 2016), a coreference corpus of Portuguese enriched with various automatic annotations.

For French however, there is no corpus of either kind, and though the reference treebank for French has been enriched with annotations for named entities (Sagot, Richard, and Stern 2012), this is merely a subset of the annotations needed for coreference.

2.2. Coreference for French and the ANCOR corpus

ANCOR is, for now, the only currently publicly available[1] large-scale coreference-annotated corpus of French, with around 418k words. It is composed of French speech transcriptions, mainly from the ESLO corpus (Baude and Dugua 2011), with coreference and morphosyntactic annotations for noun phrases and pronouns including singleton mentions, but no linguistic annotations of other elements.

More precisely, noun phrases and pronouns in ANCOR are annotated with

- Gender, number and part of speech
- Definiteness (indefinite, definite, demonstrative or expletive form)
- Inclusion or not in a prepositional phrase
- Named entity type (for named entities)
- "NEW" for the first mention of a coreference chain

[1] (Tutin et al. 2000) is another large-scale anaphora corpus but it is not publicly available.

```
<div type="section" xml:id="s2">
  <timeline>
    <when absolute="3.531" xml:id="t2.0"/>
    [...]
  </timeline>
  <u start="#t7.0" who="#spk2" xml:id="u7"
   end="#t7.19">
    [...]
    <tei:w xml:id="u7-w76">au</tei:w>
    <tei:w xml:id="u7-w77">moment</tei:w>
    <tei:w xml:id="u7-w78">où</tei:w>
    <tei:w xml:id="u7-w79">je</tei:w>
    <tei:w xml:id="u7-w80">me</tei:w>
    <tei:w xml:id="u7-w81">suis</tei:w>
    <tei:w xml:id="u7-w82">marié</tei:w>
    <tei:w xml:id="u7-w83">en</tei:w>
    <tei:w xml:id="u7-w84">juillet</tei:w>
    <tei:w xml:id="u7-w85">soixante-sept</tei:w>
  </u>
</div>
```

Figure 1: Speech transcription annotations in ANCOR-AS

```
<standOff>
  <annotation type="coreference">
    <spanGrp type="unit" subtype="mention">
      <span from="#u7-w76" to="#u7-w77"
       xml:id="m31"/>
      <span from="#u7-w84" to="#u7-w85"
       xml:id="m32"/>
    </spanGrp>
    <linkGrp type="relation" subtype="coreference">
      <link target="#m31 #m32" xml:id="r20"/>
    </linkGrp>
    <linkGrp type="schema" subtype="chain">
      <link target="#m31 #m32 #m40" xml:id="c12"/>
    </linkGrp>
  </annotation>
</standOff>
```

Figure 2: Coreference annotations in ANCOR-AS (fragments)

Another coreference corpus for French, Democrat, is under development since 2015 and is planned for release in 2019. It will include various type of documents in written French, including historic texts. The XML-TEI-URS format (Grobol, Landragin, and Heiden 2017), the basis of the format we use for ANCOR-AS, has been developed to provide a standard format in which the Democrat corpus could also be released

3. Interoperable Annotations for Coreference, Syntax and Transcription

The format of ANCOR-AS is based on the XML-TEI-URS format described in (Grobol, Landragin, and Heiden 2017) an example of which is given in fig. 1 and fig. 2.

While the original intend of XML-TEI-URS was to provide standard way of describing coreference relations, it inherits the versatility of its inspiration: the URS metamodel developed by Widlöcher and Mathet (2012) for the Glozz annotation platform. The URS metamodel provides a framework to describe relations between linguistic units with additional characterisations, making it straightforward to describe coreference relations, but can be extended to other types of annotations. In particular, describing dependency syntactic analysis is easy, since dependencies can be seen as simple typed directed relations. The XML-TEI-URS format is a TEI-compliant XML serialisation of the URS metamodel, thus enabling us to encode both coreference and syntactic annotations in a standard and interoperable way. We also chose to make those annotations stand-off, to avoid cluttering the source text too much and clashing with the speech transcription annotations. Figure 3 shows a (partial) example of syntactic annotations in XML-TEI-URS.

More precisely, a given syntactic analysis of part of an utterance in ANCOR is encoded using two layers of annotations. First, a description of the words involved using TEI `<span>` elements, that in our implementation refer to tokens `<w>` in the transcription by their `@xml:id` attribute. The `@n` attribute is also used to denote the word's index (in the Universal Dependencies sense).

The second layer is a representation of the dependency relations between words using `<link>` elements with a `@target` attribute set to `"#head_id #dependent_id"`.

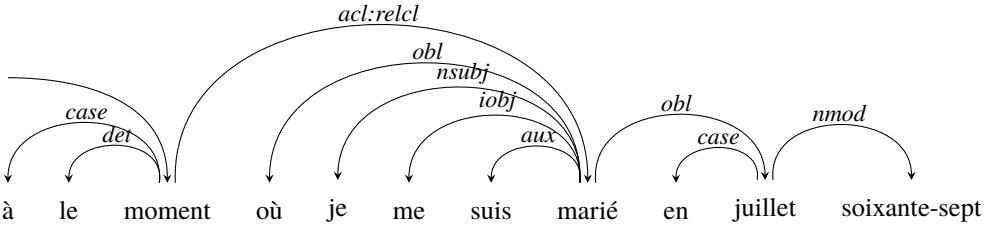
Both of theses layers are complemented by ISO feature structures (ISO 2006) describing the additional informations provided in the Universal Dependency framework, e.g. dependency types, word POS, lemma…

The only significant extension needed to adapt XML-TEI-URS to syntactic annotations was the addition of a way to describe syntactic words that are not directly accessible as a source text span. Figure 3 has an example of such an issue, as the source text token "*au*" has to be expanded into two syntactic words "*à*" and "*le*" to comply with the Universal Dependencies guidelines. These cases are dealt with by using an `<expan>` element to mark the expansion, thus providing `<w>` elements that can then be referred to in the usual way. This mirrors the recommendation of Universal Dependencies for dealing with multiword tokens. This way allows us to keep using the segmentation we use for coreference annotations to link the source text with its syntactic analysis, though it should be noted that other pointing mechanisms, such as those described in (Bański et al. 2016) might have been used, since XML-TEI-URS does not impose any particular one.

4. Enriching ANCOR with automatic syntactic annotations

For the time being, manual syntax analysis of ANCOR is out of reach, and so we settled on using automatic parsing to develop ANCOR-AS. For this iteration, we chose the DYALOG-SRNN parser (De La Clergerie, Sagot, and Seddah 2017) for its good (though unofficial) results for French in the CoNLL 2017 multilingual parsing shared task (Zeman et al. 2017). In particular, some upcoming resources for the syntactic analysis of oral French, such as those coming from the ORFEO project (Debaisieux, Benzitoun, and Deulofeu 2016), might lead us to re-evaluate our choice of parser, or at least to train it on a dataset that includes speech transcriptions.

In an effort to compensate for the fact that this parser has not been designed for speech transcriptions, we are also applying some pre-processing to the corpus in order to make it more easily parsable. Most notably, we filter out purely phatic discourse elements (such as "*hm hm*" or "*oui*"), incomplete words and fillers ("*euh*", "*ben*") before parsing (while of course preserving them in the transcription content) and segment overly-long utterances into sentences using sim-



(a) Syntactic analysis (subtree) for "*au moment où je me suis marié en juillet soixante-sept*"

```
<standOff>
  <annotation type="syntax">
    <div type="tree" xml:id="tree10">
      <div type="multiword-token">
        <expan xml:id="tree10-w6-7" n="6-7" corresp="#u7-w76">
          <w xml:id="u7-w76.1">à</w>
          <w xml:id="u7-w76.2">le</w>
        </expan>
      </div>
      <spanGrp type="unit" subtype="word">
        [...]
        <span target="#u7-w76.1" n="6" xml:id="tree10-w6" ana="#tree10-w6-fs"/>
        <span target="#u7-w76.2" n="7" xml:id="tree10-w7" ana="#tree10-w6-fs"/>
        <span target="#u7-w77" n="8" xml:id="tree10-w8" ana="#tree10-w6-fs"/>
        [...]
      </spanGrp>
      <linkGrp type="relation" subtype="dependency">
        [...]
        <link target="#tree10-w8 #tree-10-w6" xml:id="tree10-d6" ana="#tree10-d6-fs"/>
        <link target="#tree10-w8 #tree-10-w7" xml:id="tree10-d7" ana="#tree10-d7-fs"/>
        <link target="#tree10-w3 #tree-10-w8" xml:id="tree10-d8" ana="#tree10-d8-fs"/>
        [...]
      </linkGrp>
      <div type="dependency-fs">
        [...]
        <fs xml:id="tree10-d7-fs">
          <f name="type"><symbol value="det"></f>
        </fs>
        <fs xml:id="tree10-d8-fs">
          <f name="type"><symbol value="obl"></f>
        </fs>
        [...]
      </div>
      <div type="word-fs">
        [...]
        <fs xml:id="tree10-w6-fs">
          <f name="upostag"><symbol value="DET"></f>
          <f name="Definite"><symbol value="Def"></f>
          [...]
        </fs>
        [...]
      <div>
    </div>
  </annotation>
</standOff>
```

(b) Part of the corresponding XML serialisation

Figure 3: Syntactic annotations in ANCOR-AS

ple heuristics inspired by the work done for the Rhapsodie (Lacheret et al. 2014) and projects.

To ease the use of these syntactic annotations, we also provide some basic links with the coreference annotations by every mention with its syntactic head. It is obvious, though, that an automatic syntactic analysis should not be expected to be perfect, and that, in particular, the manually annotated mention spans might not match perfectly with subtrees in the syntactic analysis, as exemplified by fig. 4. In these cases, we associate mentions with the root of their minimal covering subtree, and annotate as such the dependency relations that we know to be spurious.

Though these syntactic analyses are not perfect (and unsurprisingly so, since automatic parsing of spontaneous speech is still very much an open issue), our experiments in (Grobol, Tellier, et al. 2017) give us hope that they can be of use for automatic coreference detection. Furthermore, from the perspective of the development of a real-world end-to-end coreference detection pipeline, gold-standard syntactic annotations might not be as pertinent, since such a system would still have to be able to use automatic syntactic analysis to deal with unlabeled data. Thus, while we would certainly welcome any effort of manual annotation on ANCOR, we do not consider it an absolute necessity for the avancement of automatic coreference detection for French, especially considering the recent avancements of machine learning techniques for knowledge-poor and inexact data.

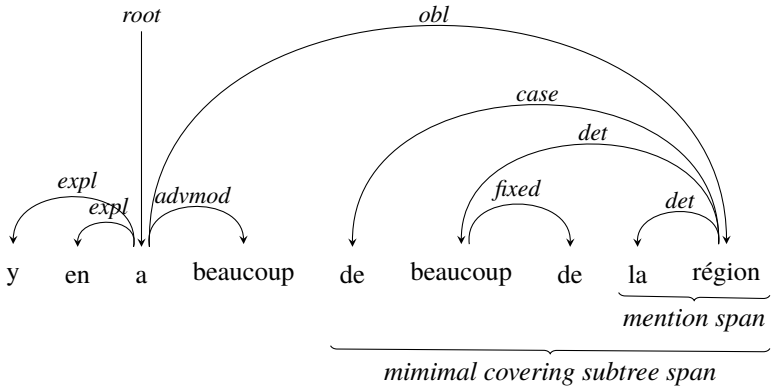


Figure 4: Bad match between syntactic analysis and mention span

5. Conclusion

In this paper, we presented an enriched version of ANCOR, which includes state-of-the-art automatic syntactic analysis and manual coreference, morphosyntactic and speech transcription annotations in a TEI-compliant format. The resulting resource is intended to serve as a stepping stone, both for the development of similar and improved coreference corpora and for the application to French of the most recent automatic coreference detection methods.

Furthermore, the specificities of coreference phenomena in spontaneous speech (such as coreferences in disfluencies, use of spatial deictics…) have not seen much interest from a corpus-based approach. We hope that providing ANCOR – one of the rare spontaneous speech corpora with coreference annotations – in an easier to use and richer format will help researchers explore this topic.

It is also our hope that this work will serve as a proof of feasibility for complex referential linguistic annotations within the TEI guidelines, at least for uses in interchange and archive formats. In this perspective, this application to dependency syntax of the XML-TEI-URS format — initially developed for coreference annotations — proves that this format is versatile enough to be used for a large class of annotation frameworks. For instance, adapting this work to add constituent-based syntactic analysis or temporal annotations (as in other ongoing projects on ANCOR) would not require significant changes to the annotation model we used here.

While further improvements to this linguistic resource are planned, the current version is available at `http://lattice.cnrs.fr/Grobol-Loic` with the same copyleft license as ANCOR (Creative Common BY-SA/BY-NC-SA).

6. Acknowledgements

This work is part of the "Investissements d'Avenir" overseen by the French National Research Agency ANR-10-LABX-0083 (Labex EFL).

This work has been supported by the ANR DEMOCRAT (Description et modélisation des chaînes de référence: outils pour l'annotation de corpus et le traitement automatique) project ANR-15-CE38-0008.

7. Bibliographical references

Antoine, J.-Y. et al. (2017). Temporal@ODIL Project: Adapting ISO-TimeML to Syntactic Treebanks for the Temporal Annotation of Spoken Speech. In H. Bunt, editor, *Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. ACL Special Interest Group on Computational Semantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2. Montpellier, France.

Bański, P. et al. (2016). *Wake up, standOff!* TEI Conference 2016. Wien, Austria.

Baude, O. and Dugua, C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*. Varia, 10: 99–118.

Calhoun, S. et al. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44.4: 387–419.

De La Clergerie, É., Sagot, B., and Seddah, D. (2017). The ParisNLP entry at the ConLL UD Shared Task 2017: A Tale of a #ParsingTragedy. In *Conference on Computational Natural Language Learning*, pages 243–252. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada.

Debaisieux, J.-M., Benzitoun, C., and Deulofeu, H.-J. (2016). Le projet ORFEO: Un corpus d'études pour le français contemporain. *Revue Corpus*. Corpus de français parlé et français parlé des corpus, 15: 91–114.

Fonseca, E. et al. (2016). Summ-it++: an Enriched Version of the Summ-it Corpus. In N. Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia. European Language Resources Association (ELRA).

Grobol, L., Landragin, F., and Heiden, S. (2017). Interoperable annotation of (co)references in the Democrat project. In H. Bunt, editor, *Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. ACL Special Interest Group on Computational Semantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2. Montpellier, France.

Grobol, L., Tellier, I., et al. (2017). Apports des analyses syntaxiques pour la détection automatique de mentions dans un corpus de français oral. In *TALN 2017*. Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Association pour le Traitement Automatique des Langues (ATALA). Orléans, France.

Hobbs, J. R. (1986). Resolving Pronoun References. In B. J. Grosz, K. Sparck-Jones, and B. L. Webber, editors, *Readings in Natural Language Processing*, pages 339–352. San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

ISO/TC 37/SC 4 (2006). *ISO 24610-1:2006 Language resource management – Feature structures – Part 1: Feature structure representation*. Reference. Geneva, CH: International Organization for Standardization.

Lacheret, A. et al. (2014). Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In *Language Resources and Evaluation Conference*. Reykjavik, Iceland.

Muzerelle, J. et al. (2014). 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 (LREC'14)*. Reykjavík, Ísland. European Language Resources Association (ELRA).

Nedoluzhko, A. et al. (2016). Coreference in Prague Czech-English Dependency Treebank. In N. Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia. European Language Resources Association (ELRA).

Nivre, J. et al. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In N. Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia. European Language Resources Association (ELRA).

Ogrodniczuk, M. et al. (2015). *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Pradhan, S., Hovy, E., et al. (2007). OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing*, pages 517–526. ICSC '07. Washington, DC, USA. IEEE Computer Society.

Pradhan, S., Moschitti, A., et al. (2012). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 1–40. CoNLL '12. Jeju, Korea. Association for Computational Linguistics.

Pradhan, S., Ramshaw, L., et al. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. CONLL Shared Task '11. Portland, Oregon. Association for Computational Linguistics.

Sagot, B., Richard, M., and Stern, R. (2012). Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In G. Antoniadis, H. Blanchon, and G. Sérasset, editors, *Traitement Automatique des Langues Naturelles (TALN)*. Volume 2 - TALN. Actes de la conférence conjointe JEP-TALN-RECITAL 2012. Grenoble, France.

Soraluze, A. et al. (2012). Mention detection: First steps in the development of a Basque coreference resolution system. In J. Jancsary, editor, *Proceedings of KONVENS 2012*, pages 128–136. Main track: oral presentations. Wien, Austria. ÖGAI.

Taulé, M., Martí, M. A., and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*. ACL Anthology Identifier: L08-1222. Marrakech, Morroco. European Language Resources Association (ELRA).

TEI consortium, editor (2016). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 3.1.0. TEI P5: Guidelines for Electronic Text Encoding and Interchange. URL: `http://www.tei-c.org/Guidelines/P5`.

Tutin, A. et al. (2000). Annotating a large corpus with anaphoric links. In *Third International Conference on Discourse Anaphora and Anaphora Resolution (DAARC2000)*, page 2. United Kingdom.

Widlöcher, A. and Mathet, Y. (2012). The Glozz Platform: A Corpus Annotation and Mining Tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, pages 171–180. DocEng '12. Paris, France. ACM.

Zeman, D. et al. (2017). CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19.