TESTACCORD Databank

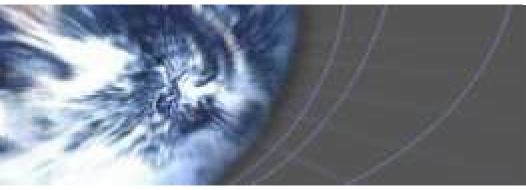
COREF data set : presentation

Jean-Yves Antoine[1], **Anaïs Lefeuvre**[1], **Jeanne Villaneau**[2]

[1]**LI – Université François Rabelais de Tours** (EA 6300)
[2]**IRISA** (UMR 6074)

Université François Rabelais Tours

http://www.info.univ-tours.fr/~antoine/parole_publique/



Introduction

TESTACCORD.COREF is a data set of *TestAccord*, a data bank of various annotation sets dedicated to the experimental study of data reliability and inter-coders agreement in the framework of Natural Language Processing.

TESTACCORD.COREF addresses the annotation of coreference and anaphora relations. More precisely, it concerns the subjective annotation of the linguistic type of previously delimitated relations. These annotation were conducted on the ANCOR_Centre coreference corpus, which can be downloaded from the Parole_Publique server too (http://www.info.univ-tours.fr/~antoine/parole_publique/).

This document presents into details the TESTACCORD.COREF data set.

TestAccord.Coref : annotated corpus

This data set results from the annotation of an restricted extract (10 files) of the ANCOR_Centre corpus, achieved by 7 different coders (version 1.0 - September 2013). The annotation of the ANCOR_Centre was conducted on the Glozz platform (Widlöcher & Mathet, 2012) and was split into three successive phases:

- Entity mentions marking,
- Referential relations marking,
- Referential relations characterization

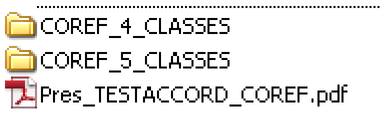
The data set TESTACCORD.COREF the phasis of characterization of the referential relations, which were already delimited. This nominal annotation consists in classifying relations among five different types:

- *Direct coreference (DIR)* – Coreferent mentions are NPs with same lexical heads.
- *Indirect coreference (IND)* – These mentions are NPs with distinct lexical heads.
- *Pronominal anaphora (PRO)* – The subsequent coreferent mention is a pronoun.
- *Bridging anaphora (BRI)* – The subsequent mention does not refer to its antecedent but depends on it for its referential interpretation(example : meronymy).
- *Bridging pronominal anaphora (BPA)* – Bridging anaphora where the subsequent mention is a pronoun. This type emphasizes metonymies (example : *Avoid Central Hostel… they are unpleasant*)

The subjects (3 men / 4 women) were adult people (average age: 43.1 years) with a high proficiency in linguistics (researchers in NLP or corpus linguistics). They know each other but worked separately during the annotation, without any restriction on time. They are considered as experts since they participated to the definition of the annotation guide. The study was conducted on an extract of 10 dialogues, representing 384 relations. This amount of annotated data satisfies the criterion of significance – at least five chance agreements per category – proposed in (Krippendorff, 2004)

TestAccord.Coref : distributed data

Two sets of test data are actually distributed, which correspond to two directories in the distribution :



COREF_4_CLASSES
COREF_5_CLASSES
Pres_TESTACCORD_COREF.pdf

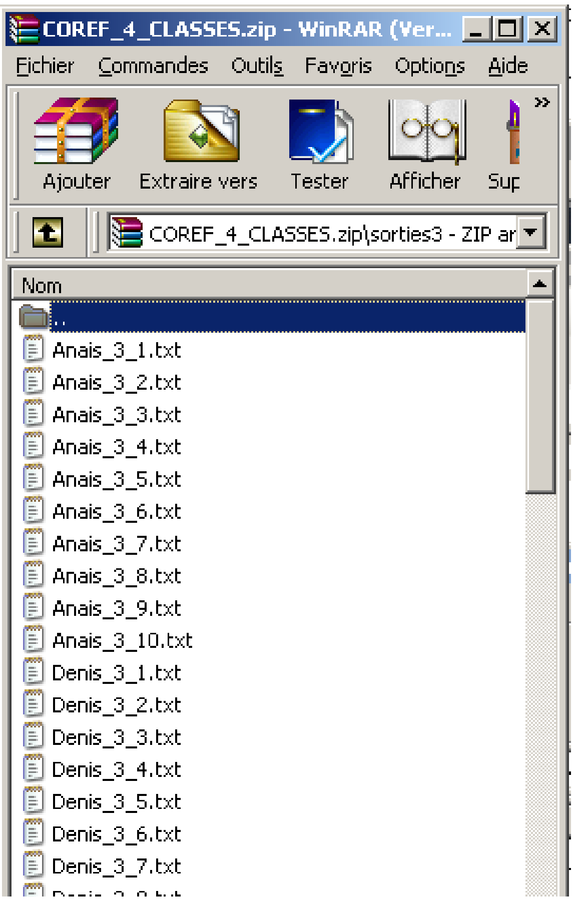
- the original 5-classes annotation (COREF_5_CLASSES directory)
- an automatically calculated 4-classes annotation, where the BRI and the IND classes are merged (COREF_4CLASSES_a) or the DIR and the IND classes are merged (COREF_4CLASSES_baseline). This merging is motivated by the potential difficulties annotators are meeting to distinguish between the BRI and IND categories (COREF_4_CLASSES_DIRECTORY)

The comparison of the two annotations will enable interested people to investigate the influence of the number of annotation classes on the same original data.

Each directory includes 70 annotated files which are corresponding to the annotations made by the seven coders on the 10 annotated files. The name of the annotated files follows the same format:

```
NameCoder_File_number.txt
```

For instance, the files `Anais_3_1.txt` and `Denis_3_1.txt` are corresponding to the annotation made respectively by the coders named *Anais* and *Denis* on the same file *3_1*. They can be compared to compute inter-coder agreement.



..
Anais_3_1.txt
Anais_3_2.txt
Anais_3_3.txt
Anais_3_4.txt
Anais_3_5.txt
Anais_3_6.txt
Anais_3_7.txt
Anais_3_8.txt
Anais_3_9.txt
Anais_3_10.txt
Denis_3_1.txt
Denis_3_2.txt
Denis_3_3.txt
Denis_3_4.txt
Denis_3_5.txt
Denis_3_6.txt
Denis_3_7.txt

Lastly, all annotated files are following the same format (see below). After a first headline, every annotation on a specific coreference relation is described by N lines (N = number of classes) of 4 columns :

- *1st column* line number
- *2nd column* relation type id (5 classes : 1 = DIR ; 2 = IND ; 3 = PRO ; 4 = BRI ; 5 = BPA)
  (4 classes : 1 = DIR ; 2 = IND ; 3 = PRO ; 4 = BRI)
- *3rd column* relation id (this id is related to the Glozz internal annotation format)
- *4th column* annotation flag (1 = the relation has been given the correspondint type ;
  0 = the relation has not been given the corresponding type

```
549Anais3.aa
1,1,jmuzerelle_1370524045887,1
2,2,jmuzerelle_1370524045887,0
3,3,jmuzerelle_1370524045887,0
4,4,jmuzerelle_1370524045887,0
5,5,jmuzerelle_1370524045887,0
6,1,jmuzerelle_1370524049622,1
7,2,jmuzerelle_1370524049622,0
8,3,jmuzerelle_1370524049622,0
9,4,jmuzerelle_1370524049622,0
10,5,jmuzerelle_1370524049622,0
11,1,jmuzerelle_1370524058470,1
12,2,jmuzerelle_1370524058470,0
13,3,jmuzerelle_1370524058470,0
14,4,jmuzerelle_1370524058470,0
15,5,jmuzerelle_1370524058470,
```

Thus, if you consider a 5-classes annotation, the annotation of a specific relation will be described by 5 successive lines, where only one of them with receive a 1 flag.

Consider for instance the following line : `1,1,jmuzerelle_1370524045887,1`. It means that the relation `jmuzerelle_1370524045887` has received the direct type (type id = 1) during the annotation.

Creative Commons Licence

TestAccord is freely distributed under a *Creative Commons* CC-BY-SA licence.

This means that you must respect the following conditions of use :

- *BY : attribution* - Licensees may copy, distribute, display and perform the work and make derivative works based on it only if they give the author or licensor the credits in the manner specified by these..

- *SA : share-alike* - Licensees may distribute derivative works only under a license identical to the license that governs the original work.

You are kindly asked to credit us by means of one of the following references:

- Muzerelle J., Lefeuvre A., Antoine J.-Y., Schang E., Maurel D., Villaneau J., Eshkol I. (2013) ANCOR : premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. Actes *TALN'2013* Les Sables d'Olonnes.
- Muzerelle J., Lefeuvre A., Schang E., Antoine J.-y., Pelletier A., Maurel D., Eshkol I., Villaneau J. (2014) ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. Proc. *LREC'2014*, Reyjavik, Iceland (submitted).

References

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

Muzerelle J., Lefeuvre A., Schang E., Antoine J.-y., Pelletier A., Maurel D., Eshkol I., Villaneau J. (2014) ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. Proc. *LREC'2014*, Reyjavik, Iceland (submitted).

Widlöcher A., Mathet Y. (2012) The Glozz platform: a corpus annotation and mining tool. *ACM Symposium on Document Engineering*. pp. 171-180