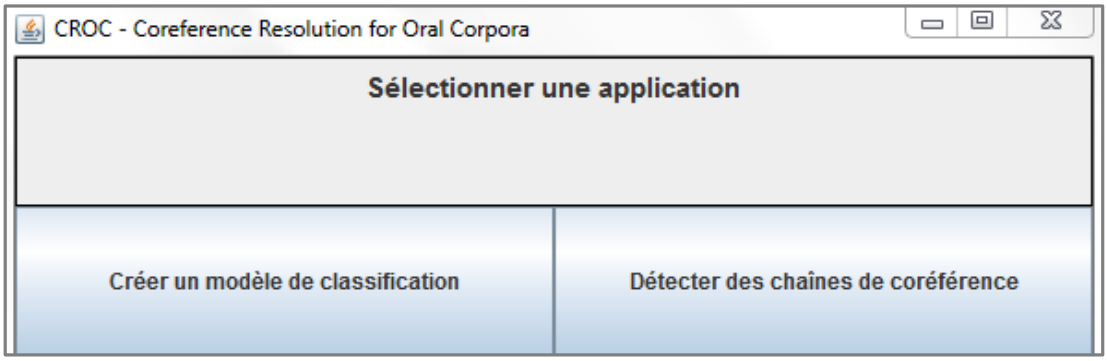
Guide d'utilisation du programme CROC
Coreference Resolver for Oral Corpora

Le programme CROC permet de détecter les chaînes de coréférence sur des données préalablement annotées en expressions référentielles (*i.e.* les unités doivent être délimitées avant d'être fournies au programme).
Le repérage des chaînes de coréférence est essentiellement basé sur une tâche de classification, pour laquelle plusieurs modèles ont été générés selon différents paramètres (corpus d'apprentissage, algorithme de classification et ensemble de traits). Afin de pouvoir mener de nouvelles expériences, le programme CROC propose également de générer de nouveaux modèles.

Ce guide détaille les deux applications proposées par CROC, soit :

1) Créer un nouveau modèle de classification
2) Annoter des données en chaînes de coréférence



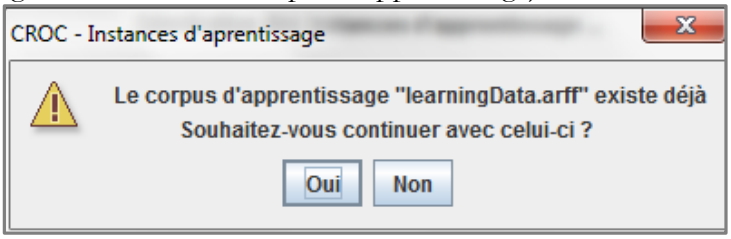
1) Créer un modèle de classification

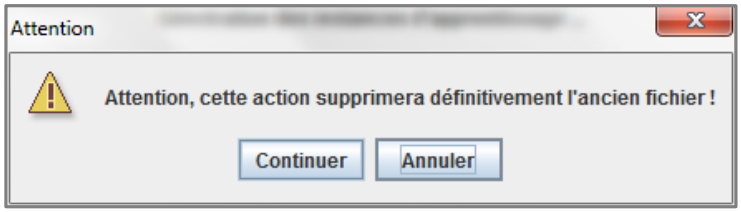
Cette tâche nécessite 3 paramètres que le programme demande à l'utilisateur :

A. Le corpus d'apprentissage
B. L'algorithme de classification
C. L'ensemble de traits

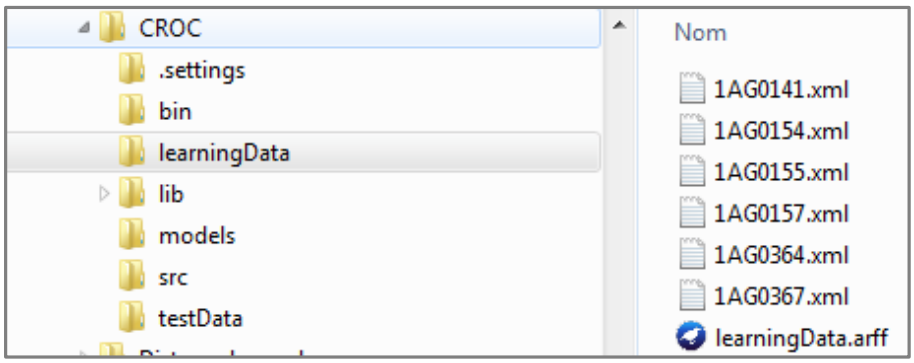
A. Le corpus d'apprentissage

Par défaut, un corpus d'apprentissage est déjà présent sous le dossier *learningData*. Il porte le nom de *learningData.arff* (c'est le fichier portant ce nom précis qui sera considéré dans la suite du programme comme le corpus d'apprentissage).





Ce fichier contient des exemples annotés qui seront nécessaires au calcul du modèle. L'utilisateur peut en créer un nouveau à partir des données de son choix, en plaçant les fichiers dont il souhaite extraires les données d'apprentissage sous le dossier *learningData*.

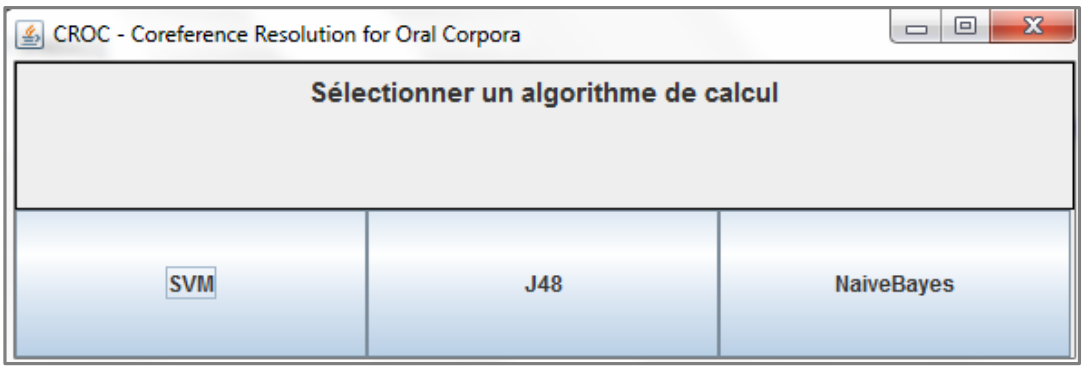


Les données des six fichiers **.xml* seront extraites pour générer un nouveau fichier *learningData.arff*

NB : Si l'utilisateur choisit de générer un nouveau corpus d'apprentissage, un certain temps est nécessaire à la création du fichier (proportionnellement au nombre de données à extraire).

B. L'algorithme de calcul

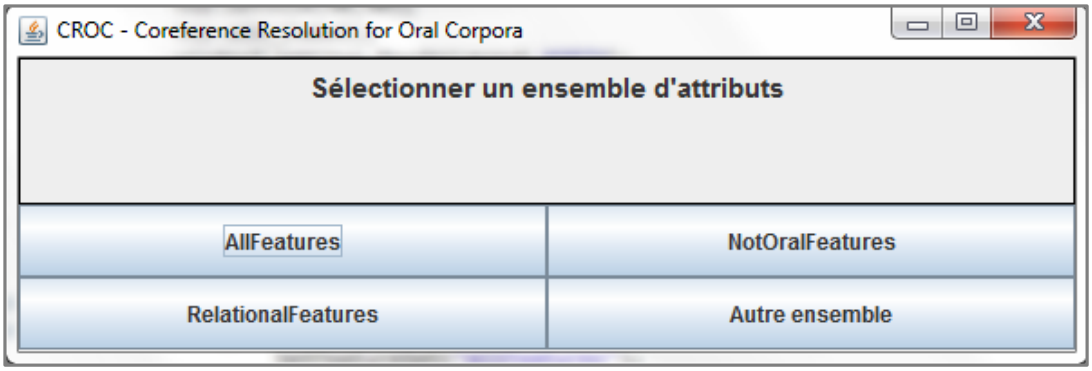
L'utilisateur a le choix entre trois algorithmes différents. Cette étape nécessite simplement un clic de l'utilisateur sur l'algorithme de son choix.



C. L'ensemble de traits

Trois ensembles d'attributs existent par défaut, et peuvent être réutilisés pour de nouveaux apprentissages :

- *allFeatureSet* : L'ensemble complet des attributs (mis à part ceux dont la valeur est une chaîne de caractères)
- *notOralFeatureSet* : Un sous-ensemble d'attributs dont ceux spécifiques à l'oral ont été retirés.
- *relationalFeatureSet* : Un sous-ensemble d'attributs comprenant uniquement les attributs relationnels (*i.e.* ceux impliquant les deux mentions d'une paire et non une mention unique)

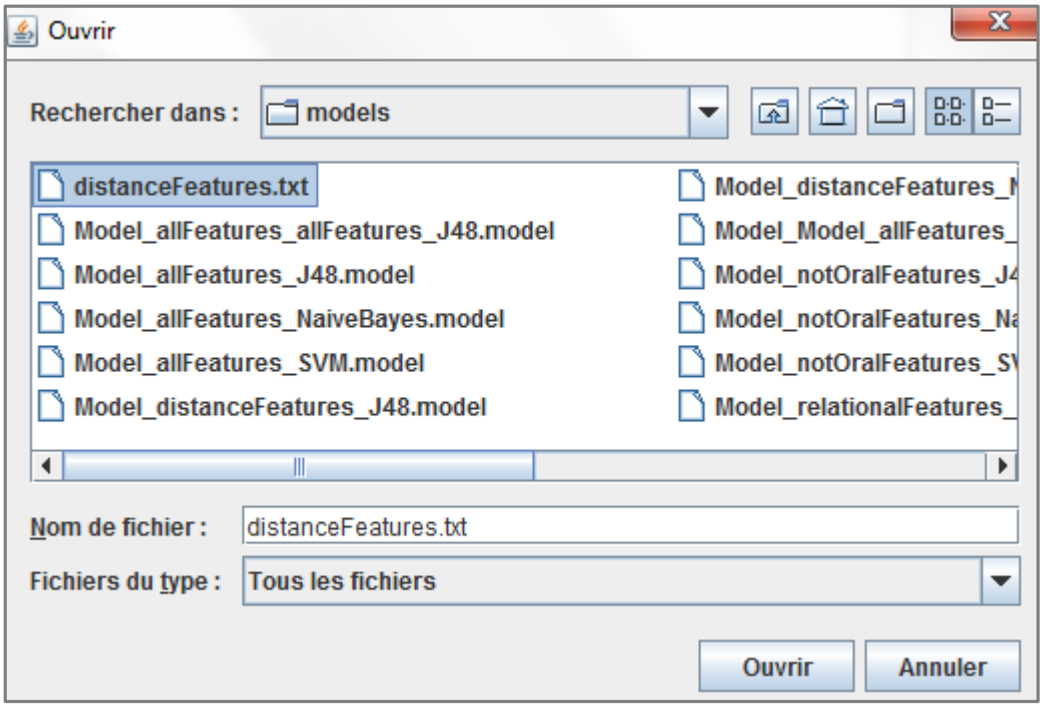


Il est également possible d'intégrer son propre ensemble de traits, en procédant comme suit :

- Composer, à l'aide de l'Annexe 1, la liste d'attributs en récupérant pour chacun son identifiant numérique unique.
- Reporter cette liste dans un fichier **.txt* sous la forme d'un identifiant par ligne
- Enregistrer le fichier sous le dossier *models* du programme (le nom attribué au fichier sera celui du nouvel ensemble d'attributs)

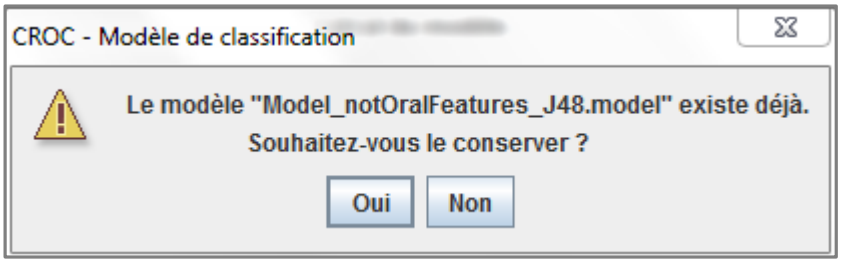
NB :

- Les attributs dont la valeur est une chaîne de caractères ne sont pas traités par le programme et seront automatiquement supprimés de l'ensemble même si l'utilisateur les sélectionne.
- Le fichier **.txt* utilisé pour créer le nouveau modèle doit être conservé puisqu'il servira à l'annotation de nouveaux fichiers selon ce modèle.
- Le nom donné au fichier **.txt* ne doit pas contenir de tiret bas



Lorsque les trois paramètres sont sélectionnés, le programme vérifie qu'un modèle généré selon ces critères n'existe pas déjà :

- Si c'est le cas, un message demande confirmation pour la suppression du précédent modèle en faveur du nouveau



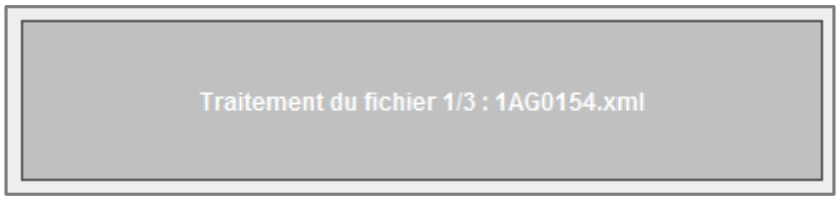
- Sinon, le modèle se calcule

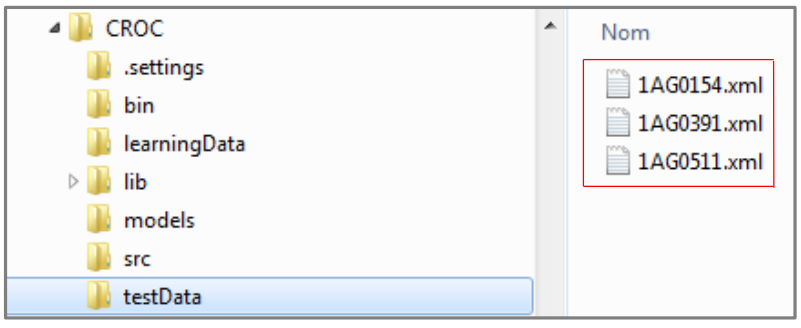
2) Détecter des chaînes de coréférence

Cette application ne nécessite aucun paramètre, mais demande en revanche de placer correctement les fichiers dans l'arborescence. En effet, pour cette étape, le programme parcours le dossier *testData*, et pour chacun des fichiers qui s'y trouve, crée un dossier *ResultFiles_** (où * correspond au nom du fichier sans son extension) pour y placer tous les fichiers crées.
En fin de traitement, un tel dossier contient :

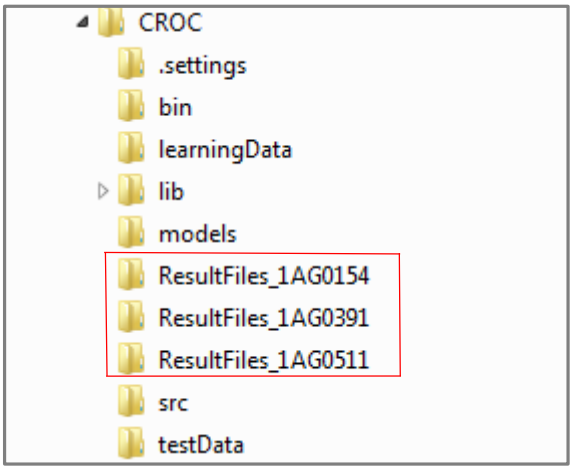
- Un fichier de référence **.arff* contenant les instances à classer avec leurs étiquettes correctes
- Un fichier ***TEI_reference_*.xml*** contenant la vraie annotation en chaînes de coréférence au format TEI (interprétable par ANALEC).
- Autant de fichiers au format TEI qu'il y a de modèles de classification.
  Ces fichiers sont de la forme ***TEI_**_system_*.xml*** (où ******** désigne le nom du modèle)
- Un fichier ***resultEval_*.txt*** contenant les résultats des différents modèles selon les 4 métriques MUC, B^3, CEAF et BLANC.
  NB : Ces résultats nécessitent la présence, dans le fichier initial, d'éléments <*relation*> grâce auxquels on déduit les vraies chaînes de coréférence à comparer aux résultats automatiques. Si ces éléments n'existent pas, le fichier de résultats n'est pas créé.

Lors du lancement de l'application, une fenêtre renseigne sur l'évolution du programme en affichant le fichier en cours de traitement.





Traitement des fichiers



ANNEXE 1 : Tableau de correspondance Attribut / Identifiant

NUMERO	ATTRIBUT
1	M2_STRING
2	M1_PREVIOUS
3	M2_NOMBRE
4	M2_EN
5	DISTANCE_MENTION
6	COM_RATE
7	M1_DEF
8	M2_PREVIOUS
9	ID_PREVIOUS
10	ID_GENRE
11	ID_FORM
12	DISTANCE_TURN
13	M2_TYPE
14	DISTANCE_WORD
15	DISTANE_CHAR
16	M2_NEXT
17	M1_FORM
18	M2_SPK
19	M2_GENRE
20	EMBEDDED
21	M1_GENRE
22	ID_SPK
23	INCL_RATE
24	ID_SUBFORM
25	ID_EN
26	M2_NEW
27	ID_NEW
28	ID_TYPE
29	M1_EN
30	ID_NEXT
31	ID_NOMBRE
32	M1_TYPE
33	M1_SPK
34	M1_NEW
35	M1_NEXT
36	M2_DEF
37	ID_DEF
38	M1_NOMBRE
39	CLASSE

NB :

- Les lignes grisées correspondent aux attributs dont la valeur est une chaîne de caractères, et ne sont pris en compte dans aucun des ensembles.
- Le dernier attribut correspondant à la classe ne peut être supprimé