Exploration du corpus ElIDA avec ScienQuest

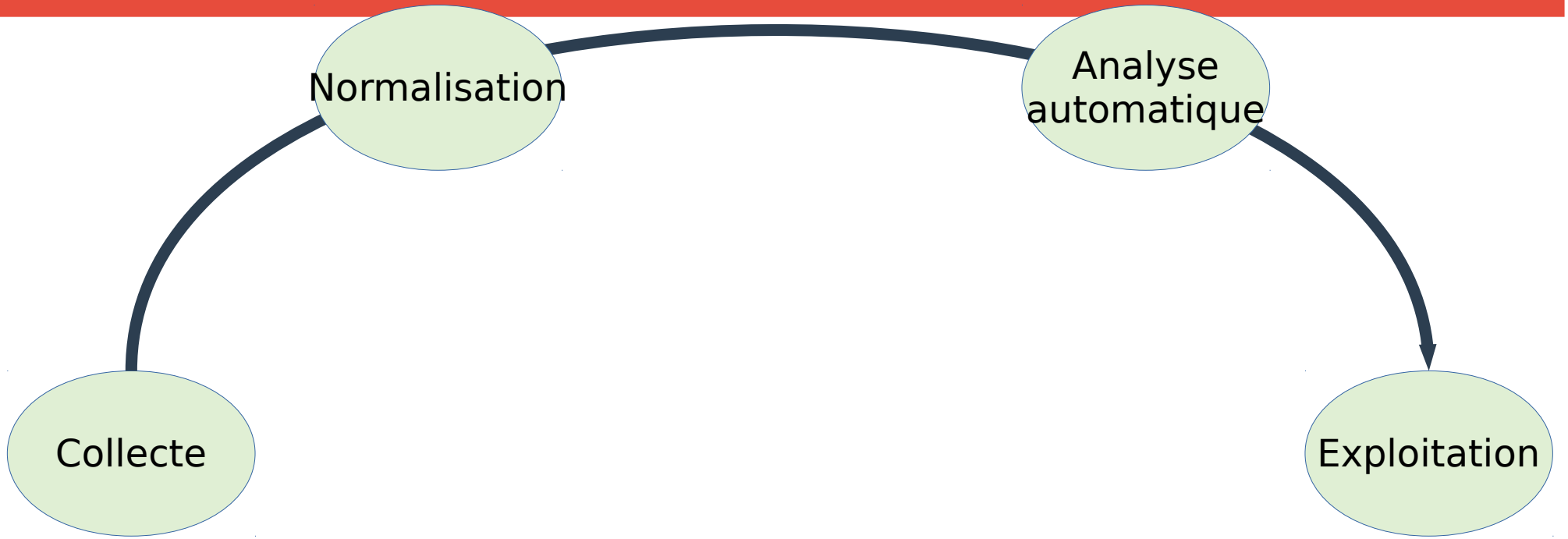
Achille Falaise
CNRS, LLF, Paris-7 Diderot

28 juin 2018
Journée d'étude *Les routines discursives dans les genres scientifiques écrits et oraux*

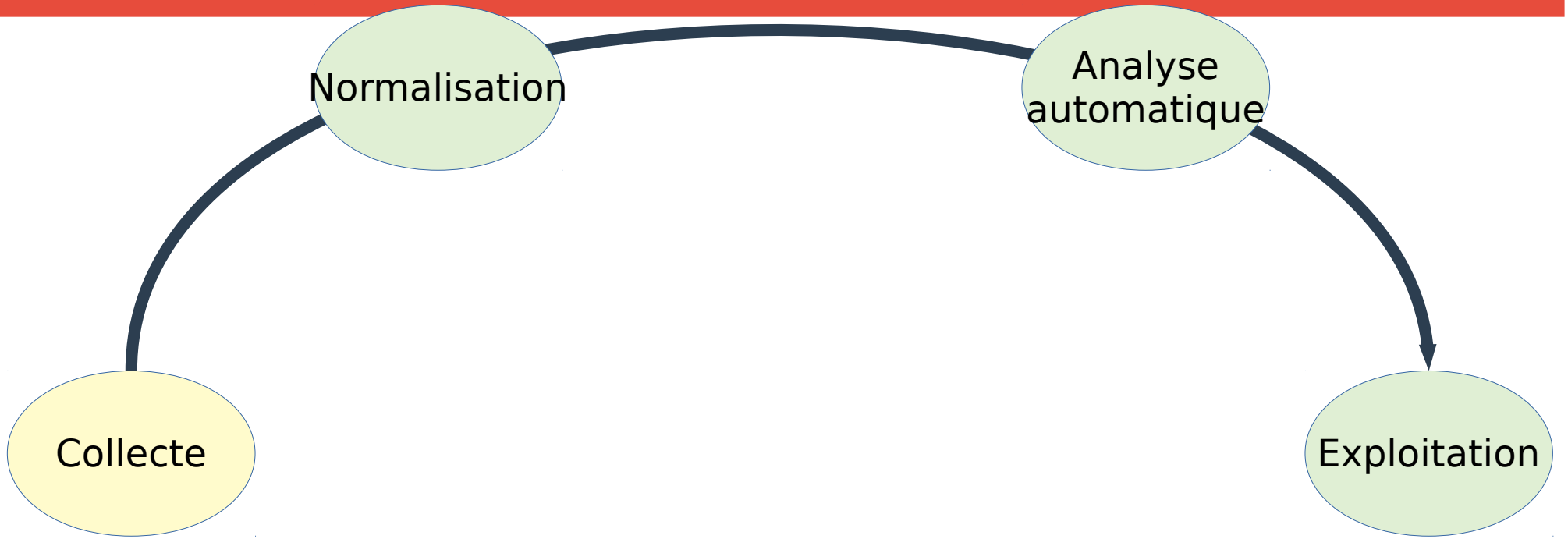
Plan

1. **Exploration de corpus**
   1. Vie d’un corpus
   2. Exploration avec ScienQUest
2. **Le corpus EIIDA**
   1. Structure du corpus
   2. Traitement du corpus
3. **Démonstration : exploration d’EIIDA avec ScienQuest**

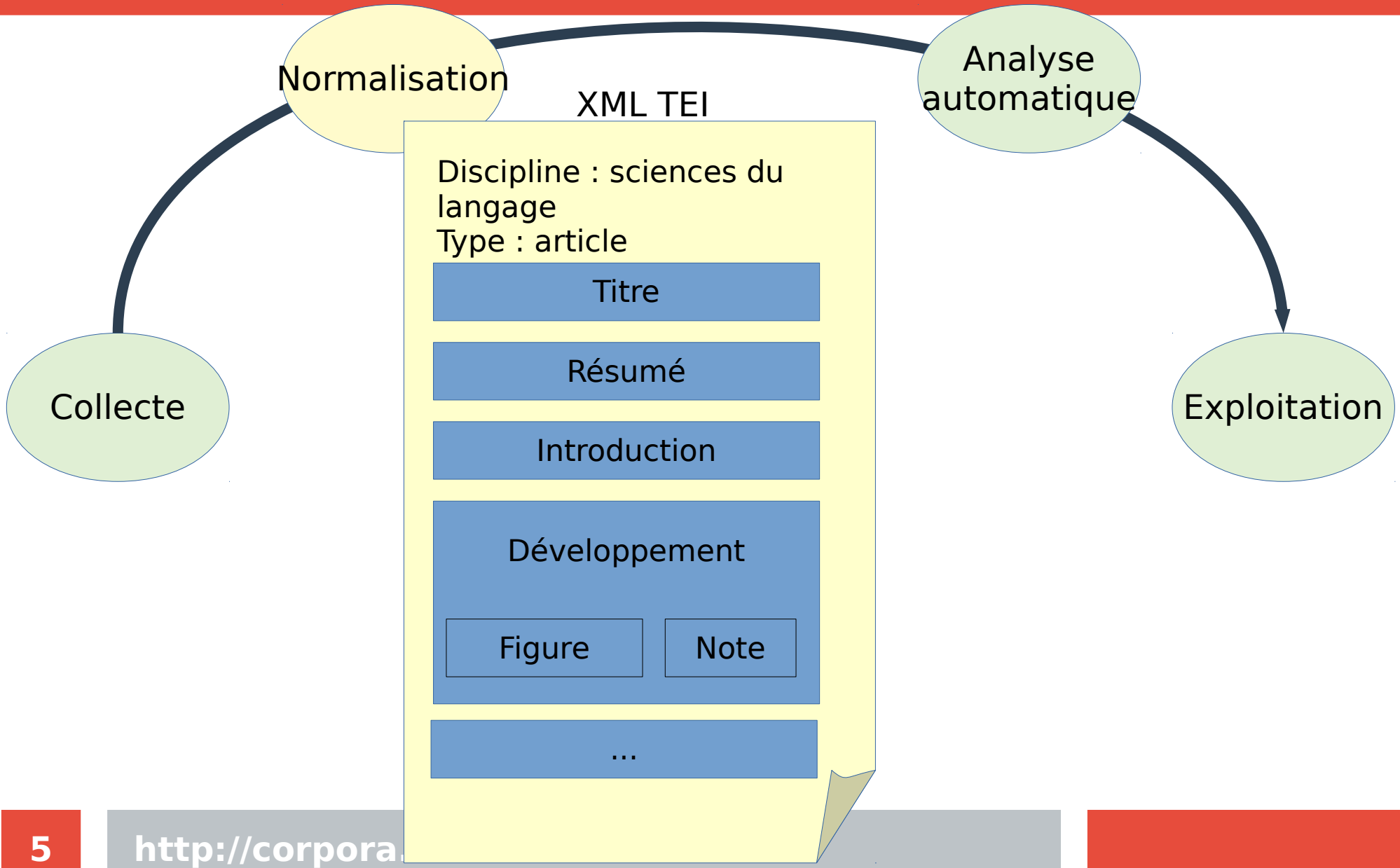
Vie d’un corpus



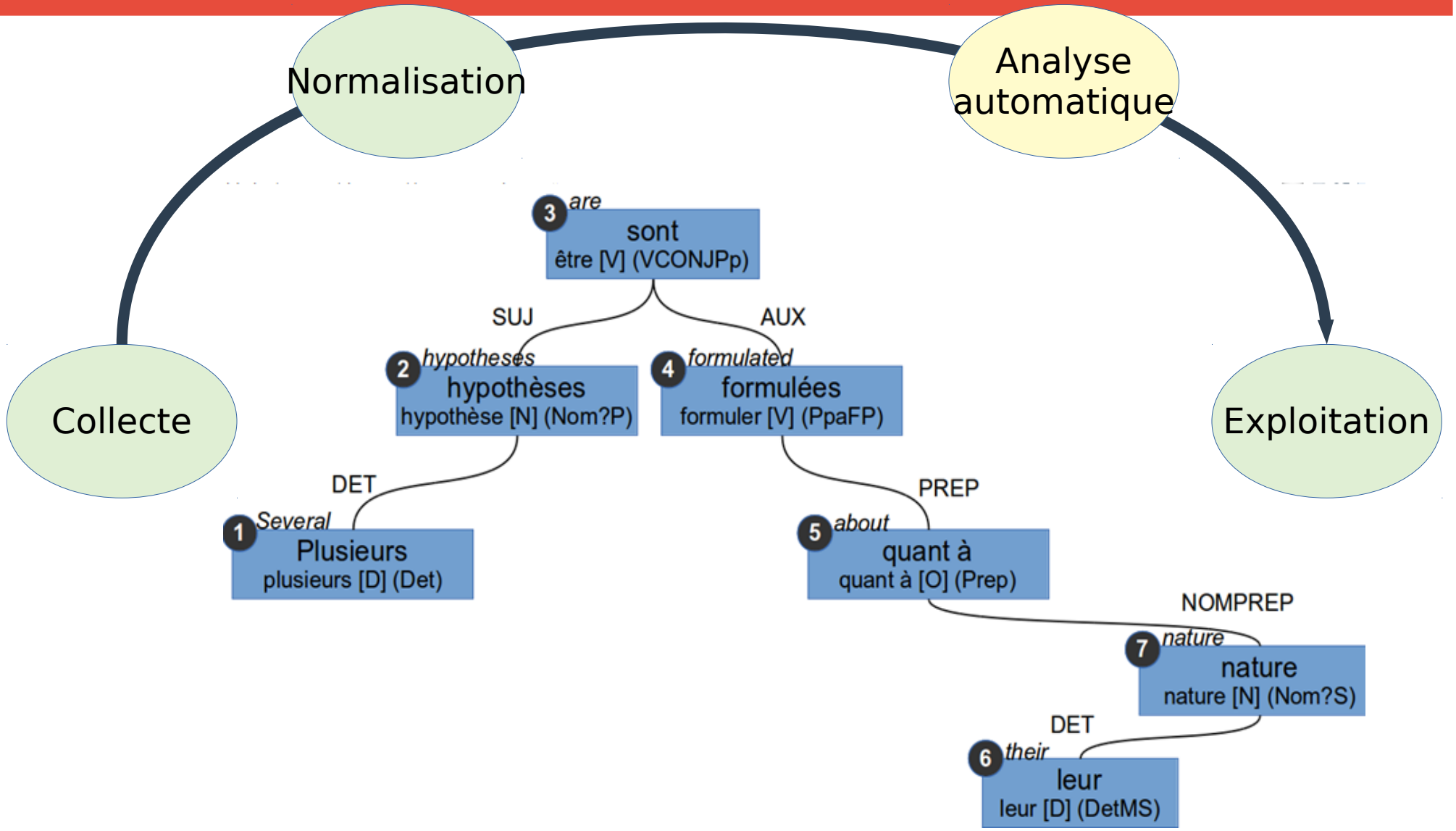
Vie d’un corpus



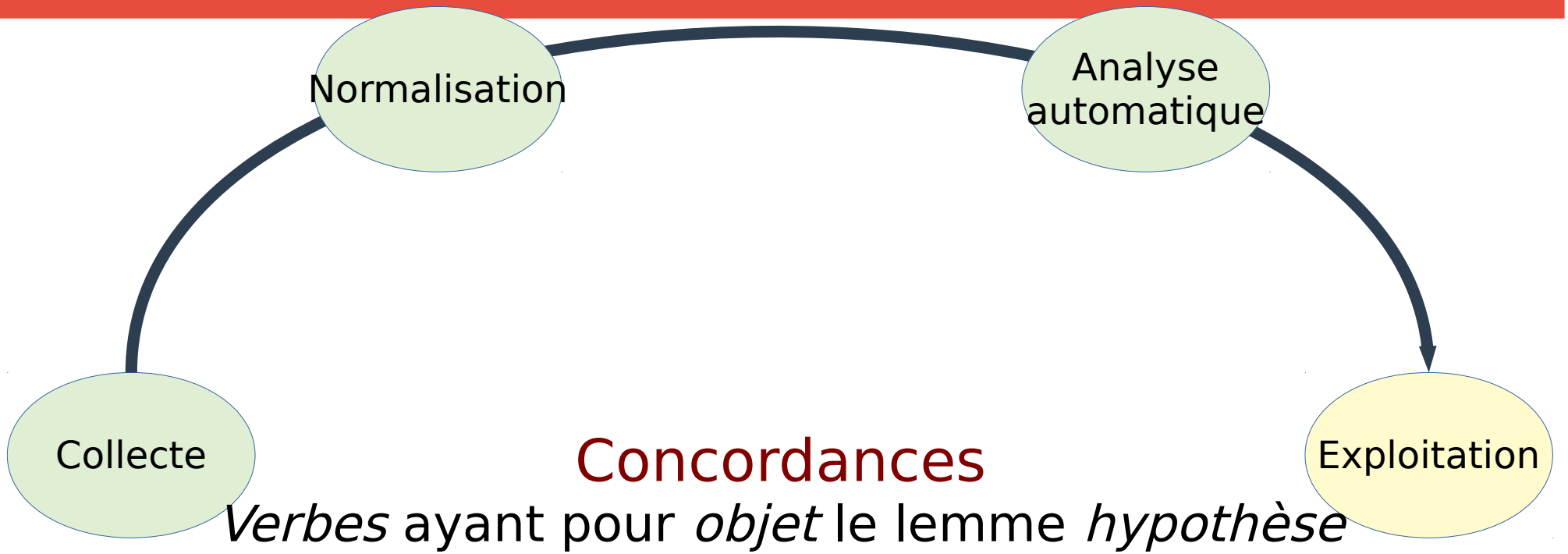
Vie d’un corpus



Vie d’un corpus



Vie d’un corpus

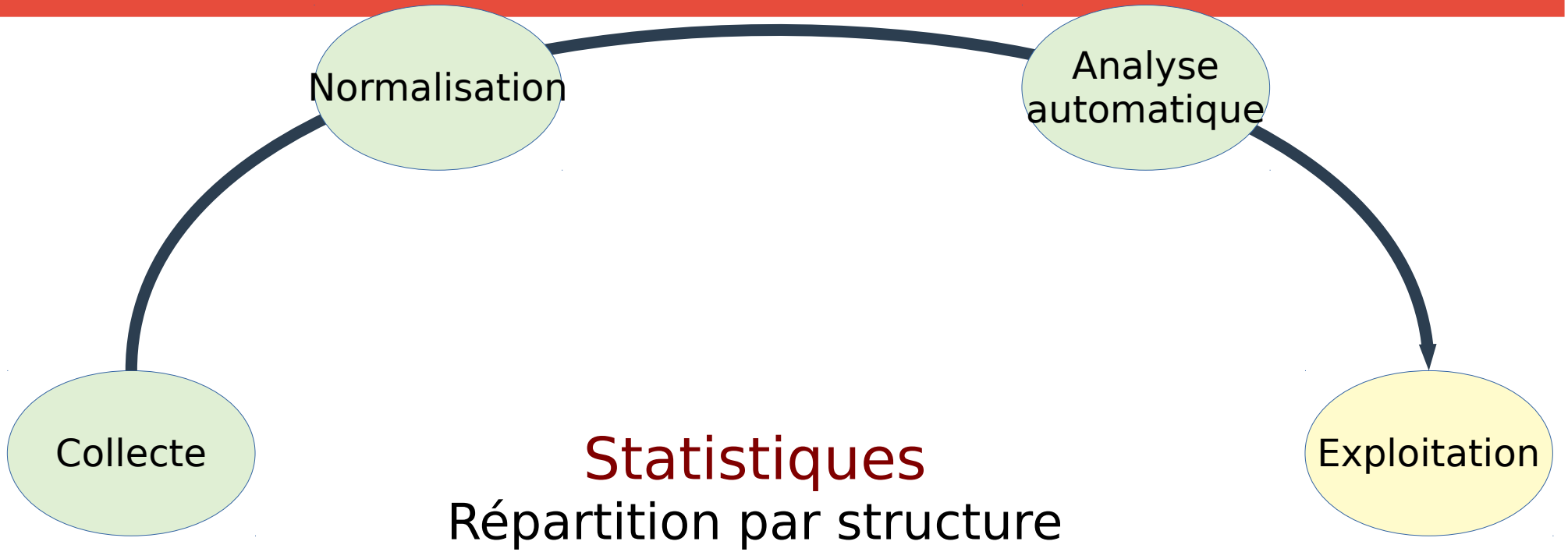


Concordances

Verbes ayant pour *objet* le lemme *hypothèse*

Contexte gauche	Occurrences	Contexte droit
démarche de décomposition des savoirs et de création de situations	implique une hypothèse	forte : c' est que l' on peut prendre ,
Il existe une réponse simple en terme de commande si l' on	fait l' hypothèse	que la perturbation (couple d' adhérence) est constante
sujets selon les types 3 et 4 , ce qui	exclut l' hypothèse	d' une détermination absolue du choix de formulation par les
marqueurs visuels : ponctuation , typographie , disposition . Nous	faisons par ailleurs l' hypothèse	que ces marqueurs risquent de varier en fonction de paramètres liés au
" , par exemple la commande Distance On retrouve avec cette	analyse l' hypothèse	issue du modèle de représentation de l' architecture textuelle ,
panoplie de moyens linguistiques pour créer du texte . Nous	faisons par ailleurs l' hypothèse	que des régularités peuvent sans doute être identifiées à condition de découper la
, à paraître : 11 - 12) . Je	formule donc sur cette base l' hypothèse	de travail selon laquelle les entités réalisées par ce que

Vie d’un corpus



Statistiques

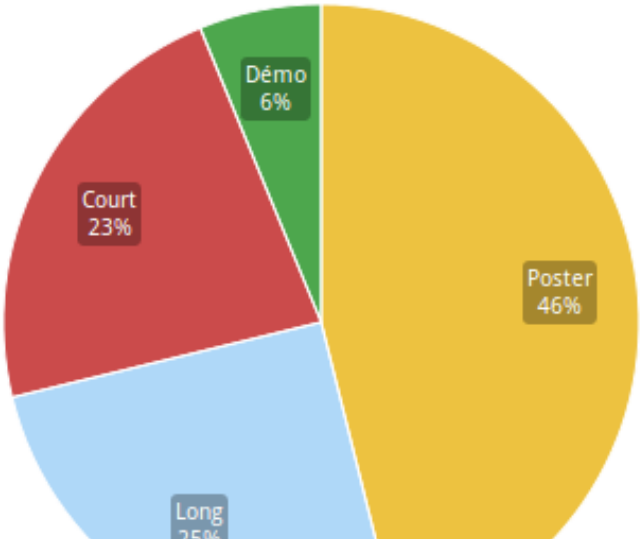
Répartition par structure

Afficher

- Forme
- Lemme
- Catégorie
- Traits morpho-syntaxiques
- Partie textuelle
- Conférence
- Type
- Année

Type

Propriété	Nb occ		Nb T		Nb norm
Long	128	/	1469430	=	0.000087
Court	52	/	663923	=	0.000078
Poster	23	/	143912	=	0.00016
Démo	1	/	46419	=	0.000022



ScienQuest

Projet ANR Scientext (2008-2009)

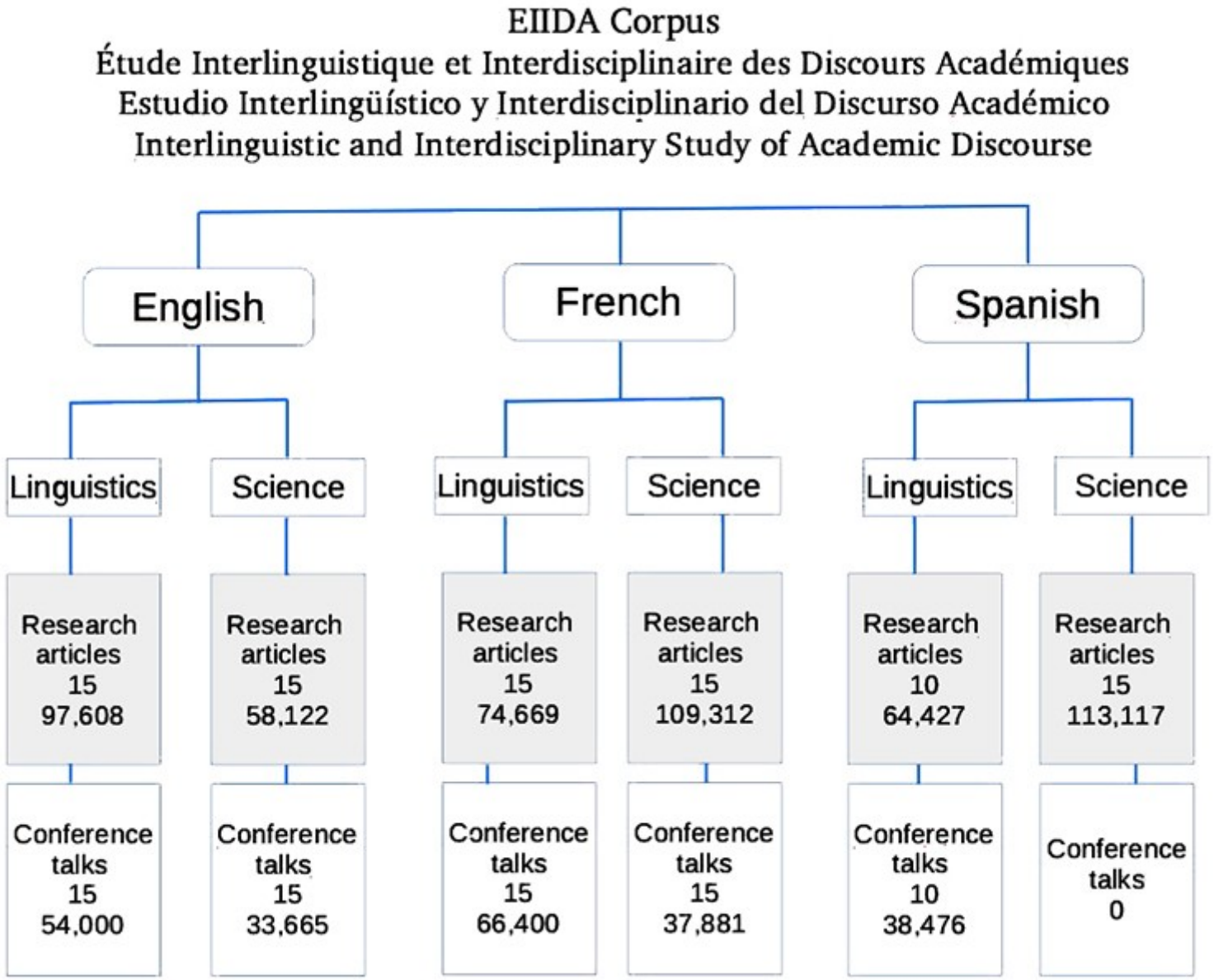
- Étude du positionnement et du raisonnement dans l’écrit scientifique
  - Phraséologie
  - Marques énonciatives
  - Marques syntaxiques
  - ... liées à la causalité
- Corpus de textes scientifiques
  - Français (5M mots, 8 disciplines)
  - Anglais (14M mots, biologie + médecine)
  - Textes argumentatifs d'apprenants de l'anglais (1M mots)
  - Évaluations d'articles (502 évals)

Corpus EllDA

Études interdisciplinaires et interlinguistiques du discours académique (2012-2017)

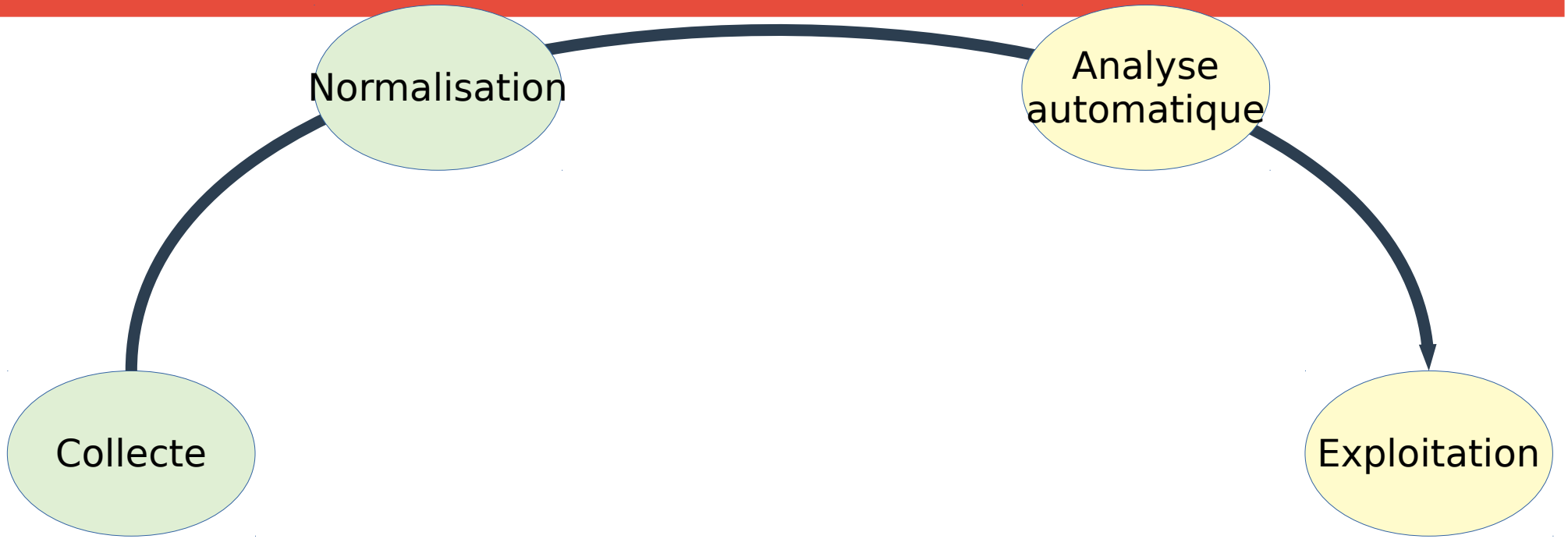
- Corpus comparable
  - Comparaison modalité (écrit / parole)
  - Comparaison langue (anglais / espagnol / français)
- Un corpus trilingue ; pour chaque langue :
  - 60 textes
  - 30 communications orales (transcrites et relues – pour l'anglais et le français) + 30 publications écrites
  - 2 disciplines (géochimie et linguistique)
- Total : 300k mots (parole, soit ~ 20h) + 900k mots (écrit)
- Plus d'infos : http://www.transfers.ens.fr

Corpus EIIDA

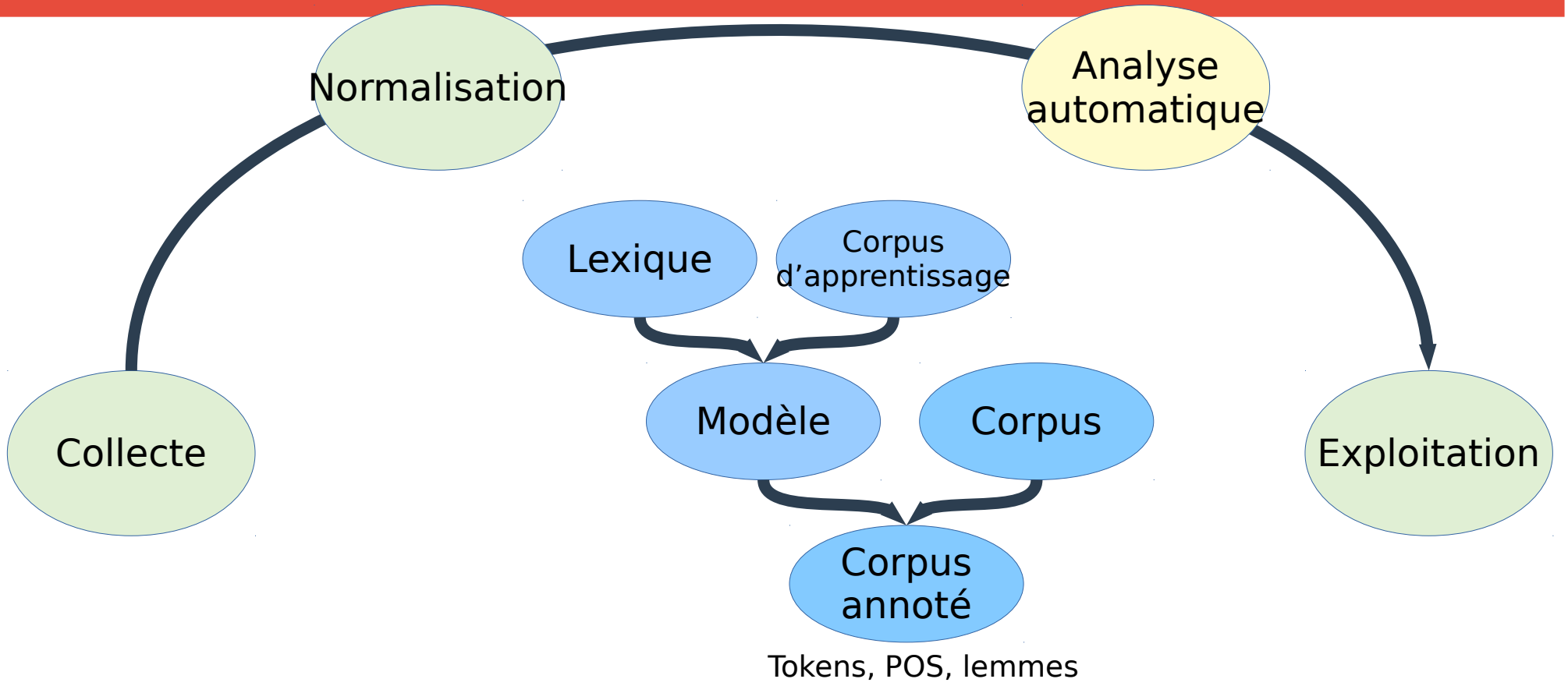


http://www.transfers.ens.fr

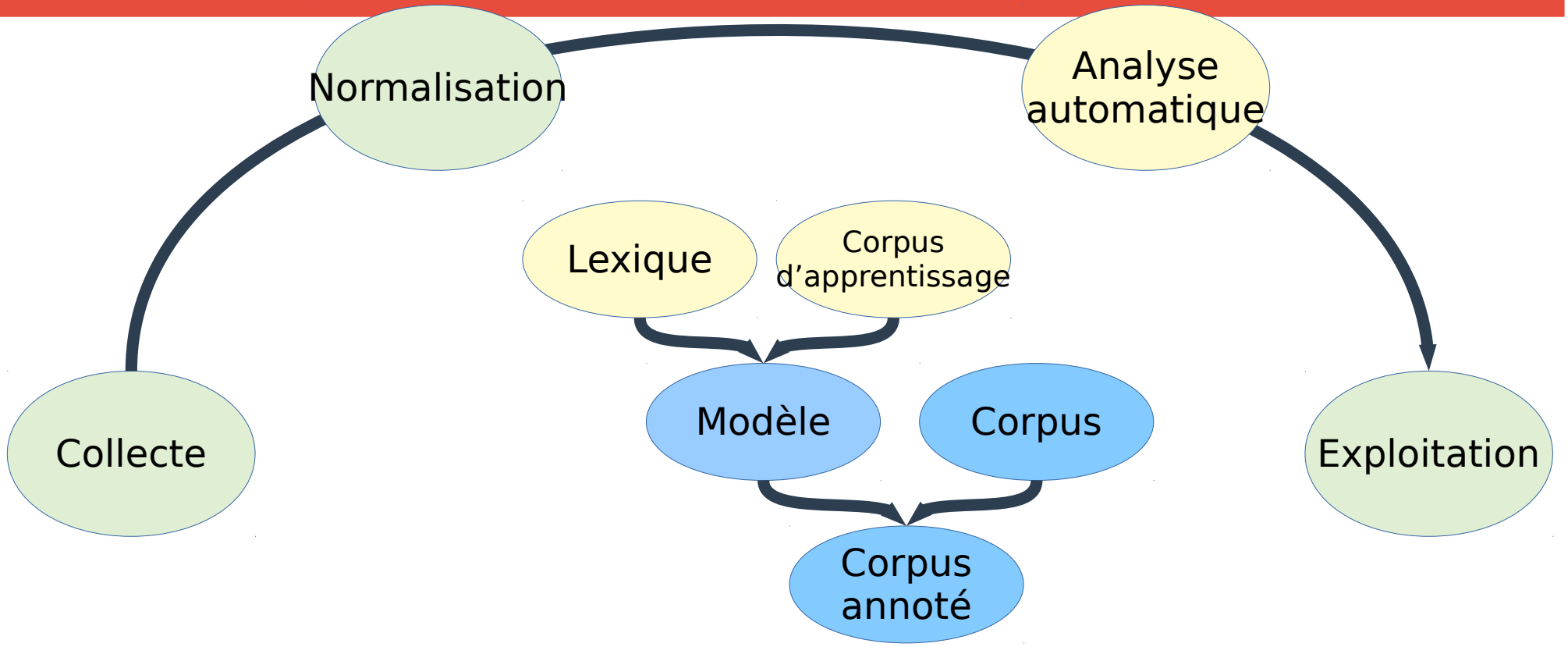
Corpus EllDA



Corpus EllDA

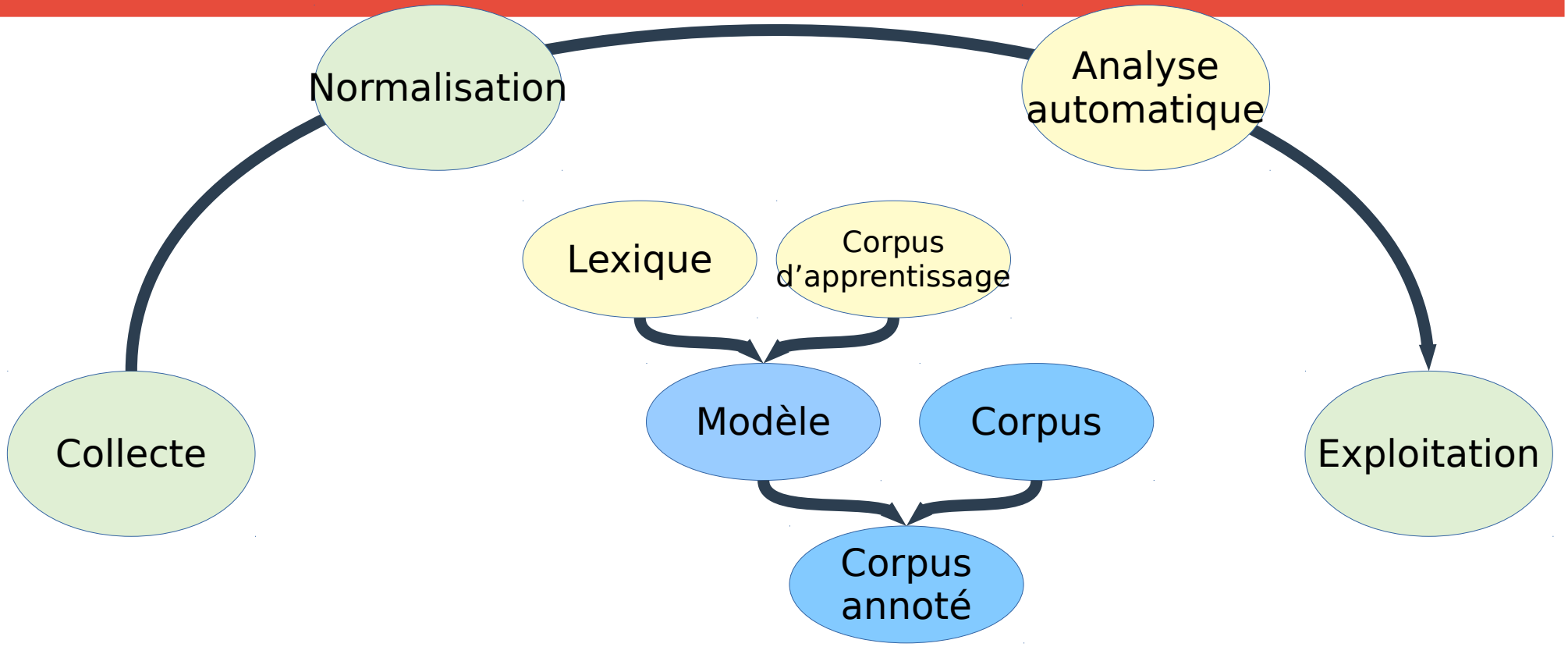


Corpus EllDA



Détermine :
- type de texte
- jeu d'étiquettes
- tokenisation

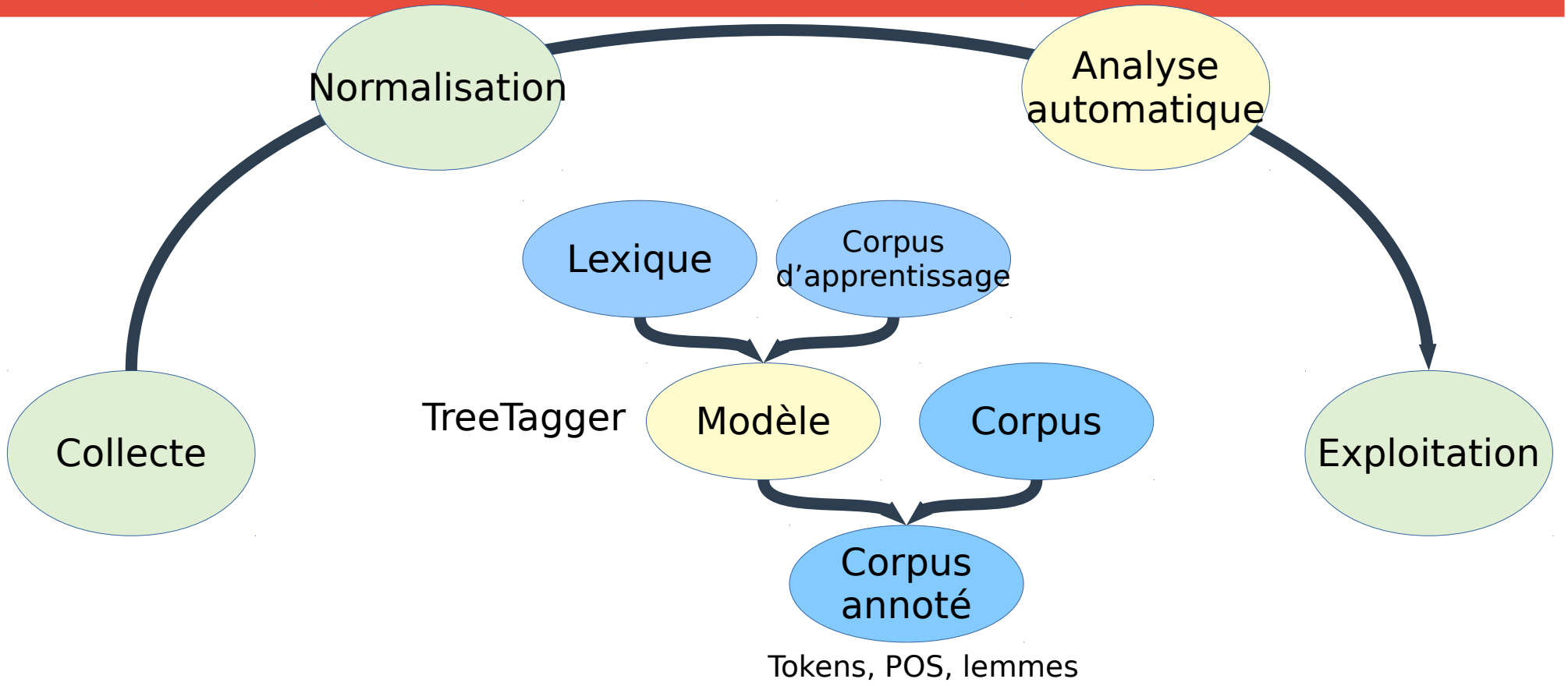
Corpus ElIDA



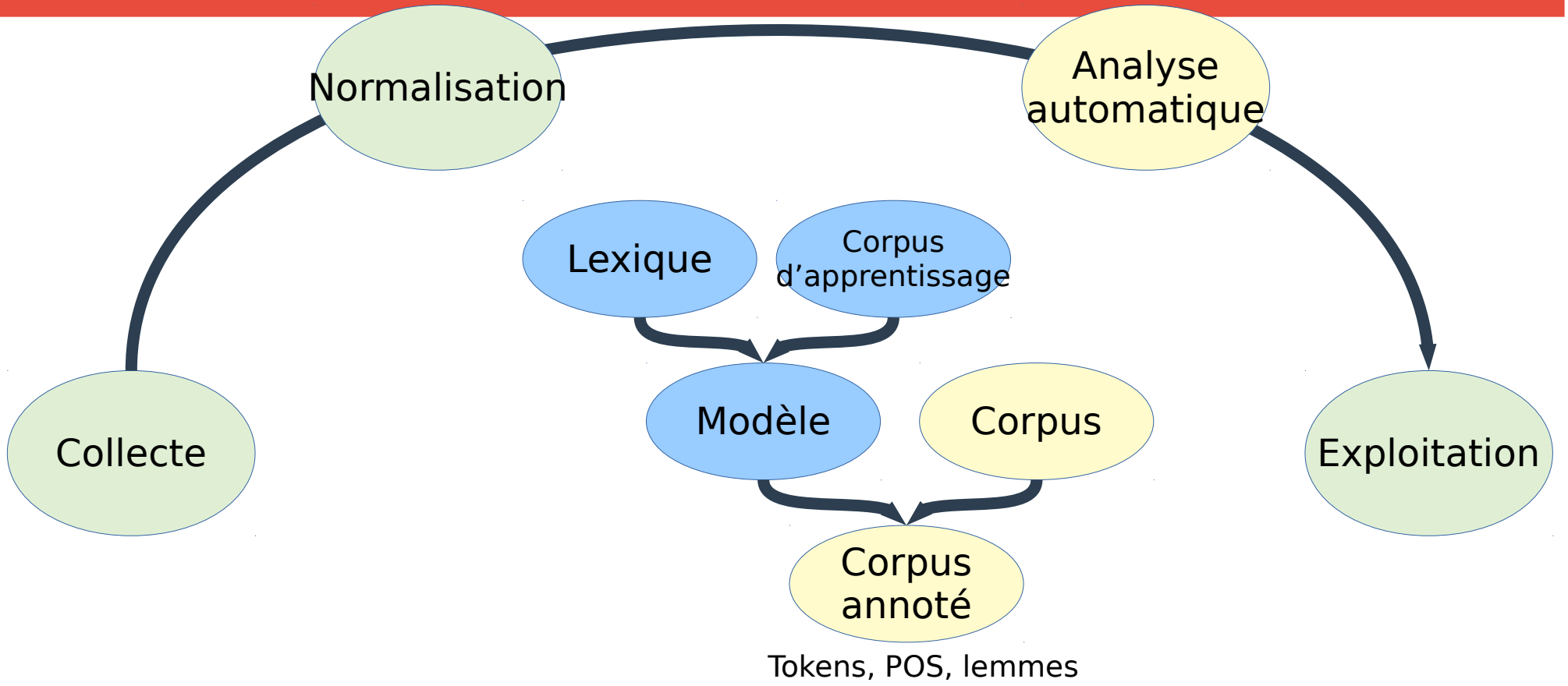
Détermine :
- type de texte
- jeu d'étiquettes
- tokenisation

Parce que
En fait
Par conséquent

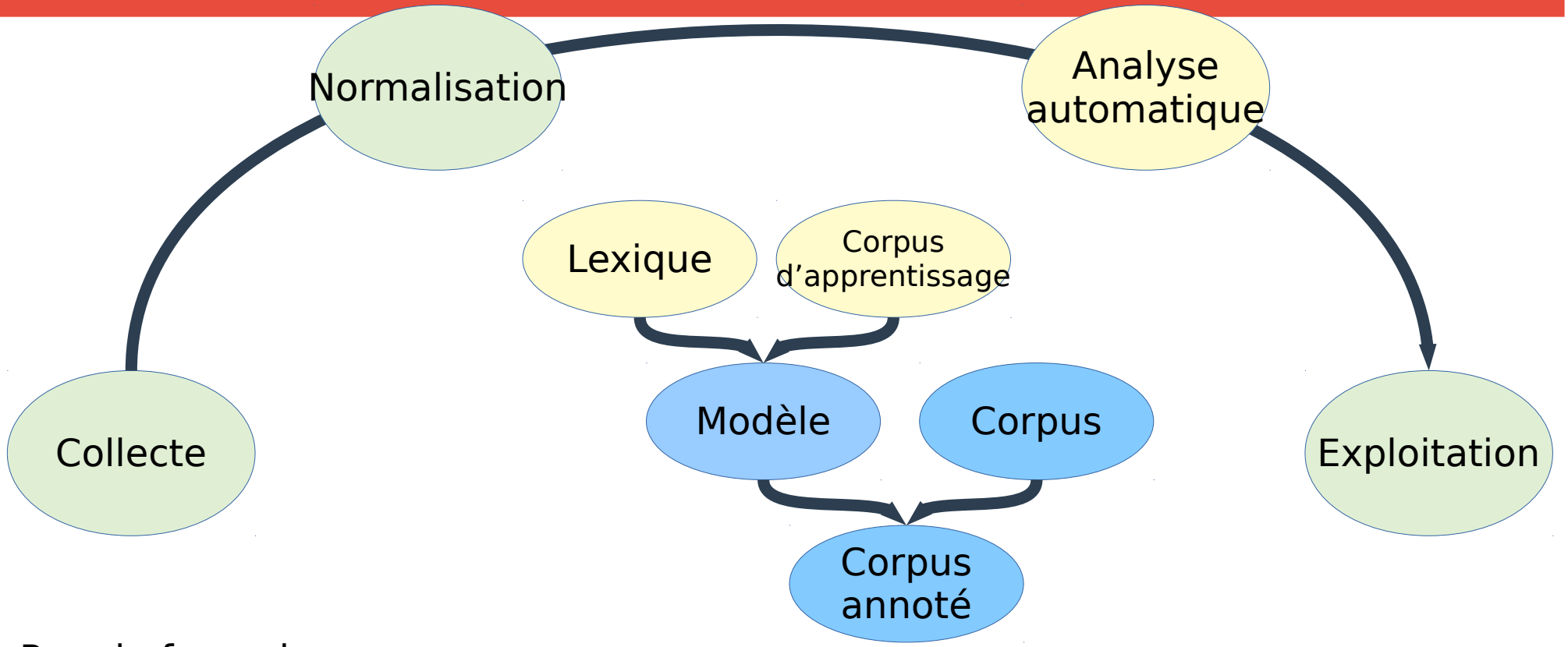
Corpus EllDA



Corpus ElIDA



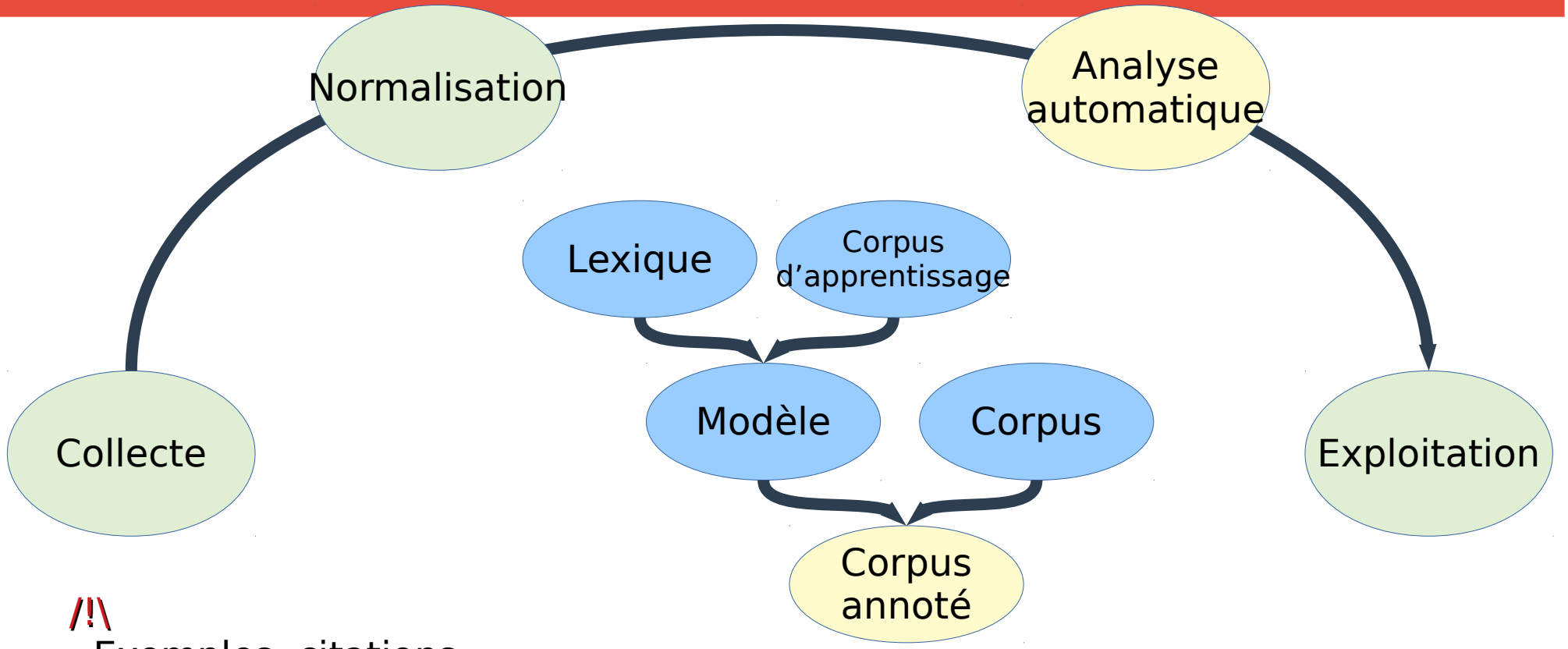
Corpus ElIDA



Pour le français :
- corpus et jeu d'étiquettes PERCEO

Pour l'anglais :
- corpus SUZANNE, jeu d'étiquettes ~Lancaster

Corpus ElIDA



/!\
Exemples, citations
Découpage en phrases
Locuteurs

Démo : exploration d'EIIDA avec ScienQuest

Quelques locutions adverbiales

En fait

Quand même

Quelles sont les locutions adverbiales les plus fréquentes ?

Séquences d'adverbes

ADV^n

Emploi de *On*

On peut VER

On VER VER:Inf

Pour le corpus anglais :
Login : eiida
Mdp : eiida