Achille Falaise

Corpora for human beings

LingLunch, 21 juin 2018

Achille Falaise

Corpora for human beings

LingLunch, 21 ju

- linguists
- translators
- language teachers
- language learners
- foreign language speakers
- tourists
- computer scientists

Achille Falaise

Corpora for human beings

Lir and by in 2018

Achille Falaise

Corpora for human beings

Lin

and by

As long as it's written :

- Computer chat,
- Audio transcripts,
- Unformal/popular/old/weird writing

... are welcome !

1) Background

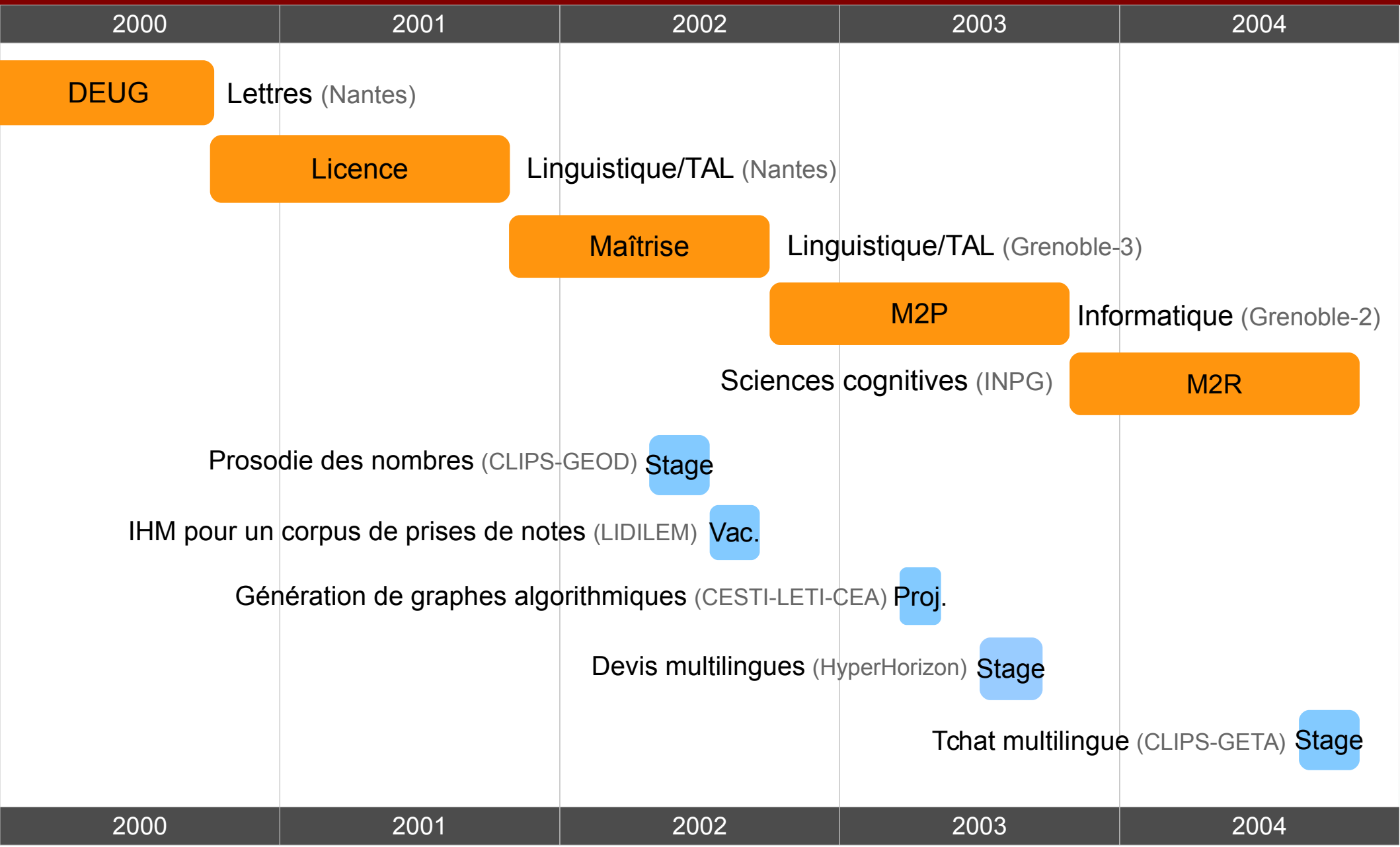
2) Past work

- Internet Relay Chat (*corpora by humans*)
- Scientific texts (*corpora for humans*)
- Corpora for teaching (*corpora for humans*)
- « Old » French texts (*corpora by humans*)

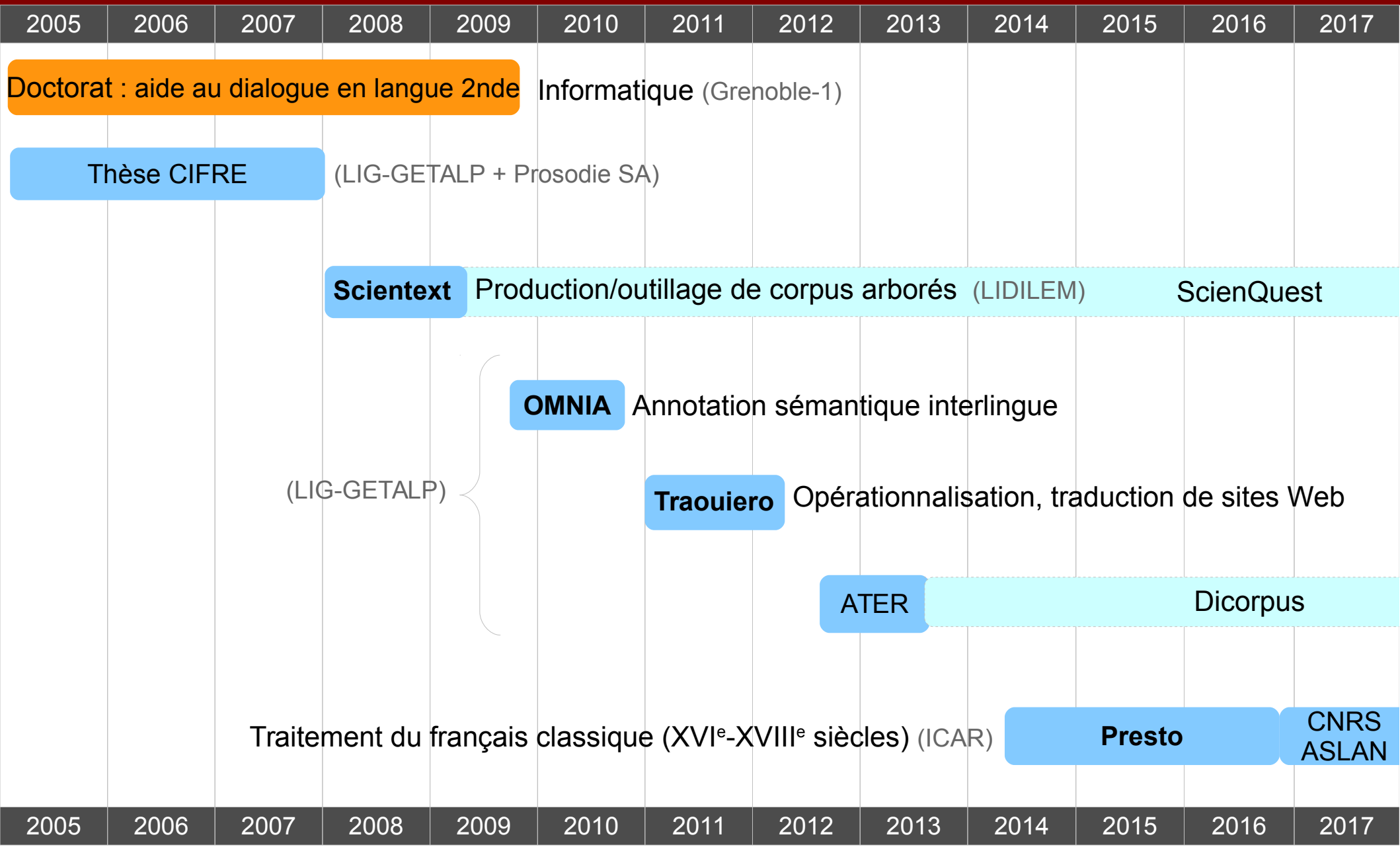
3) Future work

- Self-made corpora for everyone

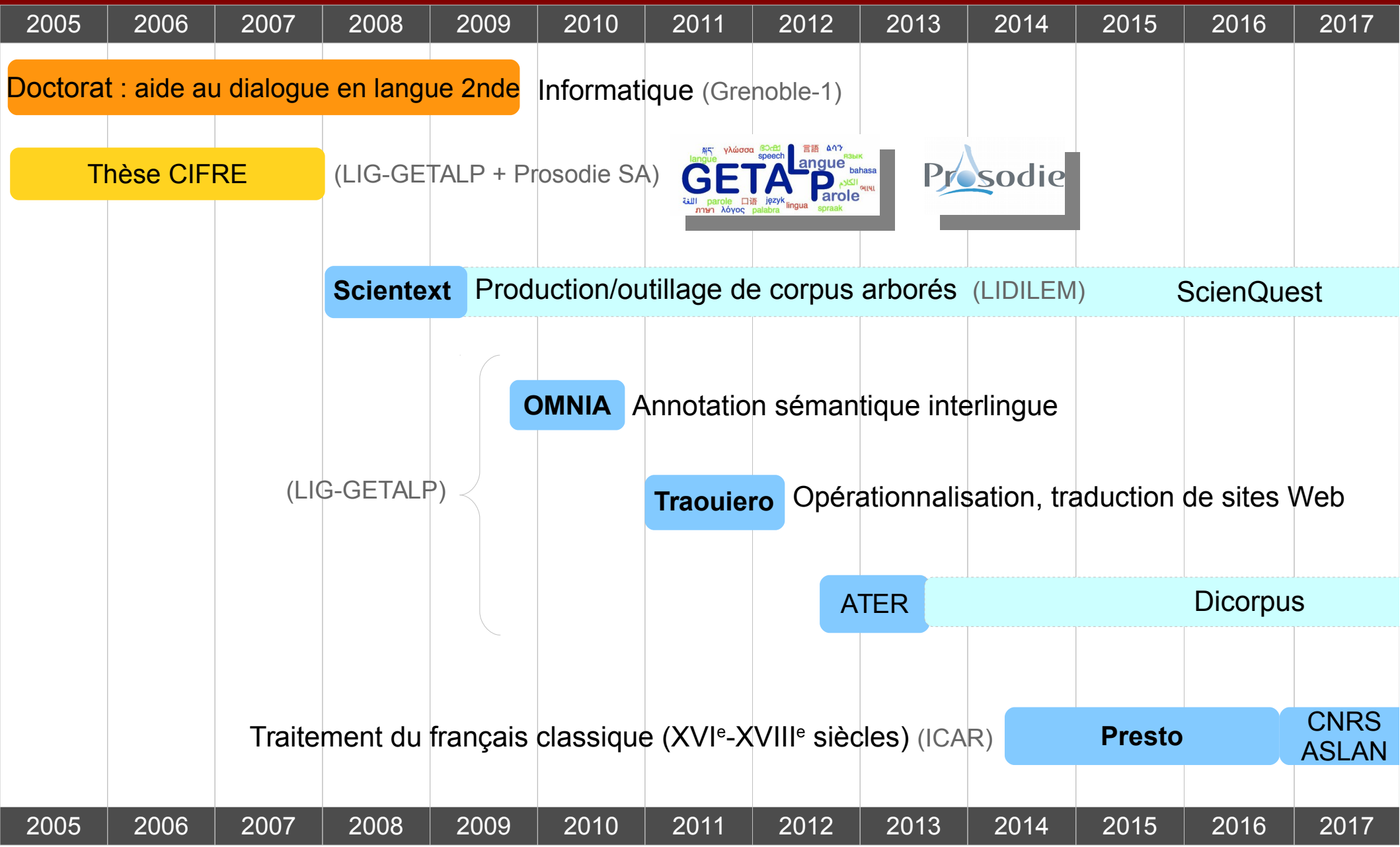
Background


2000
2001
2002
2003
2004
DEUG
Lettres (Nantes)
Licence
Linguistique/TAL (Nantes)
Maîtrise
Linguistique/TAL (Grenoble-3)
M2P
Informatique (Grenoble-2)
Sciences cognitives (INPG)
M2R
Prosodie des nombres (CLIPS-GEOD) Stage
IHM pour un corpus de prises de notes (LIDILEM) Vac.
Génération de graphes algorithmiques (CESTI-LETI-CEA) Proj.
Devis multilingues (HyperHorizon) Stage
Tchat multilingue (CLIPS-GETA) Stage
2000
2001
2002
2003
2004

Background


2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Doctorat : aide au dialogue en langue 2nde
Informatique (Grenoble-1)
Thèse CIFRE
(LIG-GETALP + Prosodie SA)
Scientext
Production/outillage de corpus arborés (LIDILEM)
ScienQuest
OMNIA
Annotation sémantique interlingue
(LIG-GETALP)
Traouiero
Opérationnalisation, traduction de sites Web
ATER
Dicorpus
Traitement du français classique (XVIe-XVIIIe siècles) (ICAR)
Presto
CNRS
ASLAN

Background


2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Doctorat : aide au dialogue en langue 2nde
Informatique (Grenoble-1)
Thèse CIFRE
(LIG-GETALP + Prosodie SA)
Scientext
Production/outillage de corpus arborés (LIDILEM)
ScienQuest
OMNIA
Annotation sémantique interlingue
(LIG-GETALP)
Traouiero
Opérationnalisation, traduction de sites Web
ATER
Dicorpus
Traitement du français classique (XVIe-XVIIIe siècles) (ICAR)
Presto
CNRS
ASLAN

M2 Cognitive Science internship, LIG-GETALP

Multilingual chat

- Creating a (demo) translating service for chat : easy
- Making it usefull : a tougher nut to crack

2004 : Can Internet chat actually be translated ?

- « It's too noisy »
- « It's youth language, only them can understand »
- « It's just (badly) transcripted speech »
- « It's text with noise added to look like speech »

Corpus

- Collection from free (of charge) access IRC logs, no licence
- 4M messages, 23M tokens, 105 chanels, 3 months, CC-BY-NC licence

« Corpus du français tchaté »

Home]

Canal
18-25ans

Nb d'interv. / page
500

Interventions
- Messages
- Commandes
- Evénements

Valider

Page de résultats
[1] 1 à 500

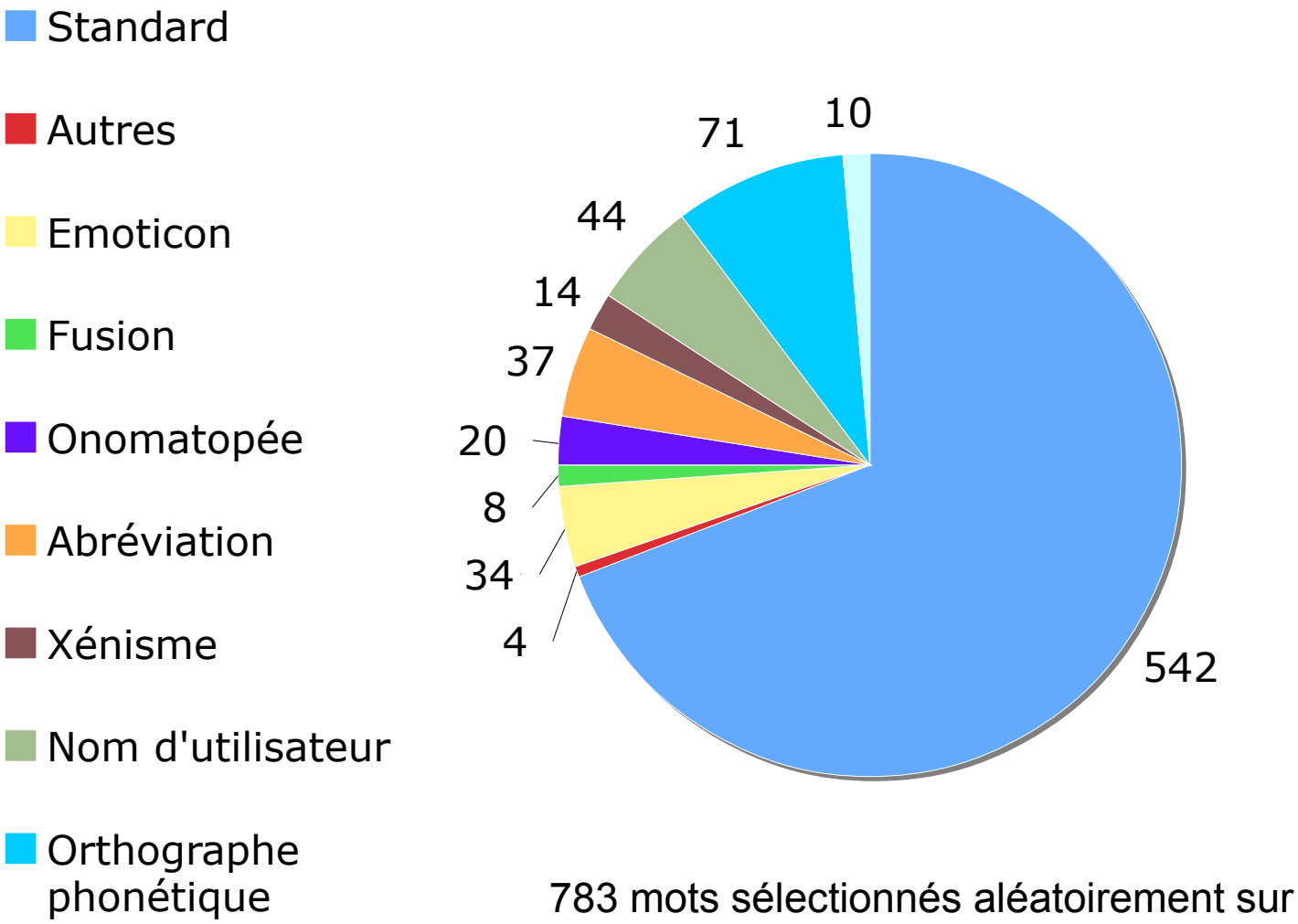
Regex
/ / OK

Infos sur la sélection [+]
527290 interventions.
2788893 mots.

Canal *18-25ans*

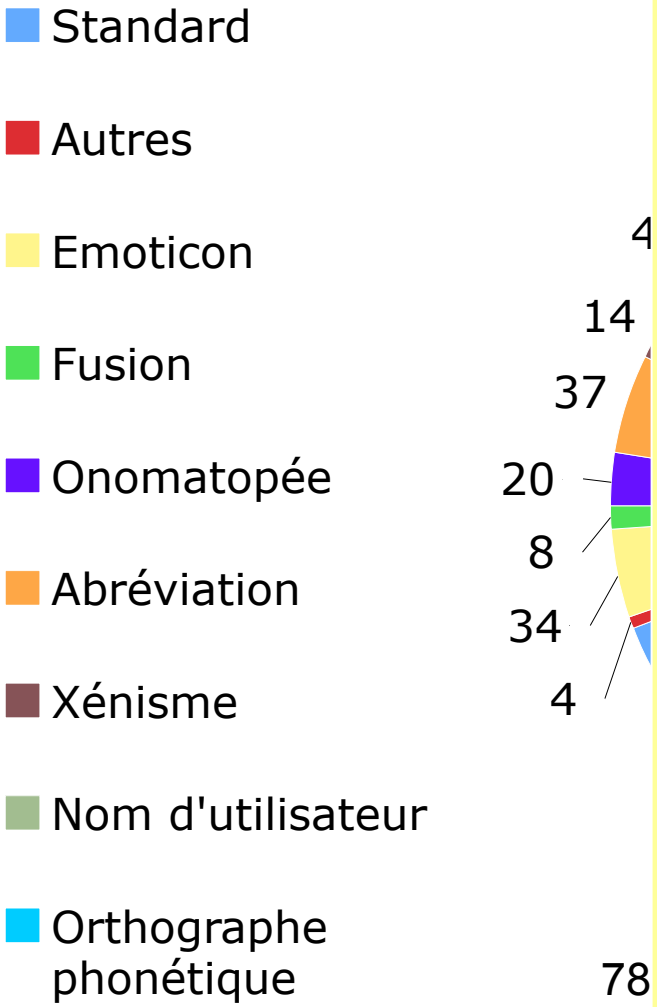
N°Int	Date	Heure	N°Aut.	Nom/Type	Texte
1	7/1/2004	0:0	Prue	Message	alalala alors
2	7/1/2004	0:0	DataStats02	Message	.:: Les statistiques de ce canal sont disponibles sur: http://go.botstats.com/?18-25ans ::
3	7/1/2004	0:0	Sniper_	Message	re Melanie
4	7/1/2004	0:0		Événement	(~aaron@EpHslo5by1.fs) s'est déconnecté: Quit: Quit
5	7/1/2004	0:0		Événement	ShadowTiger is listening to LES WRIGGLES - La pâte à modeler ` R-script ,
6	7/1/2004	0:0		Événement	Changement de pseudo: Camyre|away -> Camyre|boulot
7	7/1/2004	0:0	Prue	Message	je m'atta au pire
8	7/1/2004	0:0	Prue	Message	ptdr
9	7/1/2004	0:1	Prue	Message	j'avais raison de m'attendre au pire Bwet_La_Mwete ^^
10	7/1/2004	0:1	Bwet_La_Mwete	Message	un peu de respect.
11	7/1/2004	0:1	Bwet_La_Mwete	Message	=)
12	7/1/2004	0:1	Prue	Message	c'est toi ?
13	7/1/2004	0:1	Prue	Message	euh
14	7/1/2004	0:1	Prue	Message	de toi ?
15	7/1/2004	0:1	Shaolan	Message	Bonne nuit tout le monde (Survivor 4 Episode 3 Powa)
16	7/1/2004	0:1	Bwet_La_Mwete	Message	oué
17	7/1/2004	0:1	Bwet_La_Mwete	Message	xD
18	7/1/2004	0:1	Prue	Message	oué à laquelle ? :p
19	7/1/2004	0:1	Prue	Message	bn Shaolan
20	7/1/2004	0:1	Bwet_La_Mwete	Message	oué au deux
21	7/1/2004	0:1	Bwet_La_Mwete	Message	=)
22	7/1/2004	0:1	Prue	Message	ah ok
23	7/1/2004	0:1	Prue	Message	un auto portrait donc
24	7/1/2004	0:1	Bwet_La_Mwete	Message	tu remarquera le beau sceptre du pepsi =)

Aperçu du français tchaté



783 mots sélectionnés aléatoirement sur le canal #18-25ans

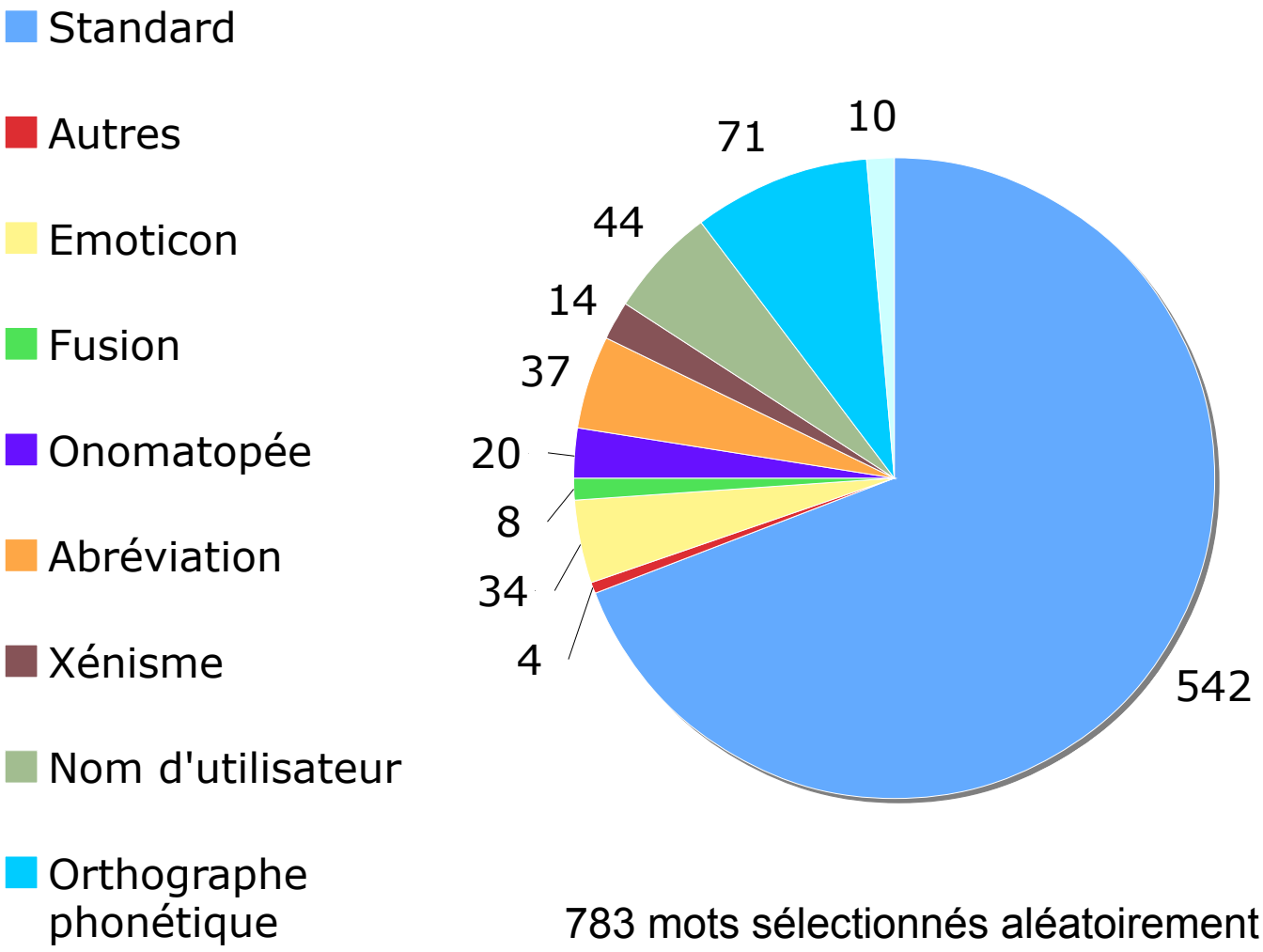
Aperçu du français tchaté



Orthographe phonétique

A> soir tlm
A> kikooooooooo *B* :*)
B> kikooo *A* :)
<*C* vient de se connecter>
A> yo *C*
C> yop all
C> yop *A*
A> o y repond today
C> kikouuu *B*
C> yop midam *B*
C> yop *D*
A> j'ai de la chance loll

Aperçu du français tchaté



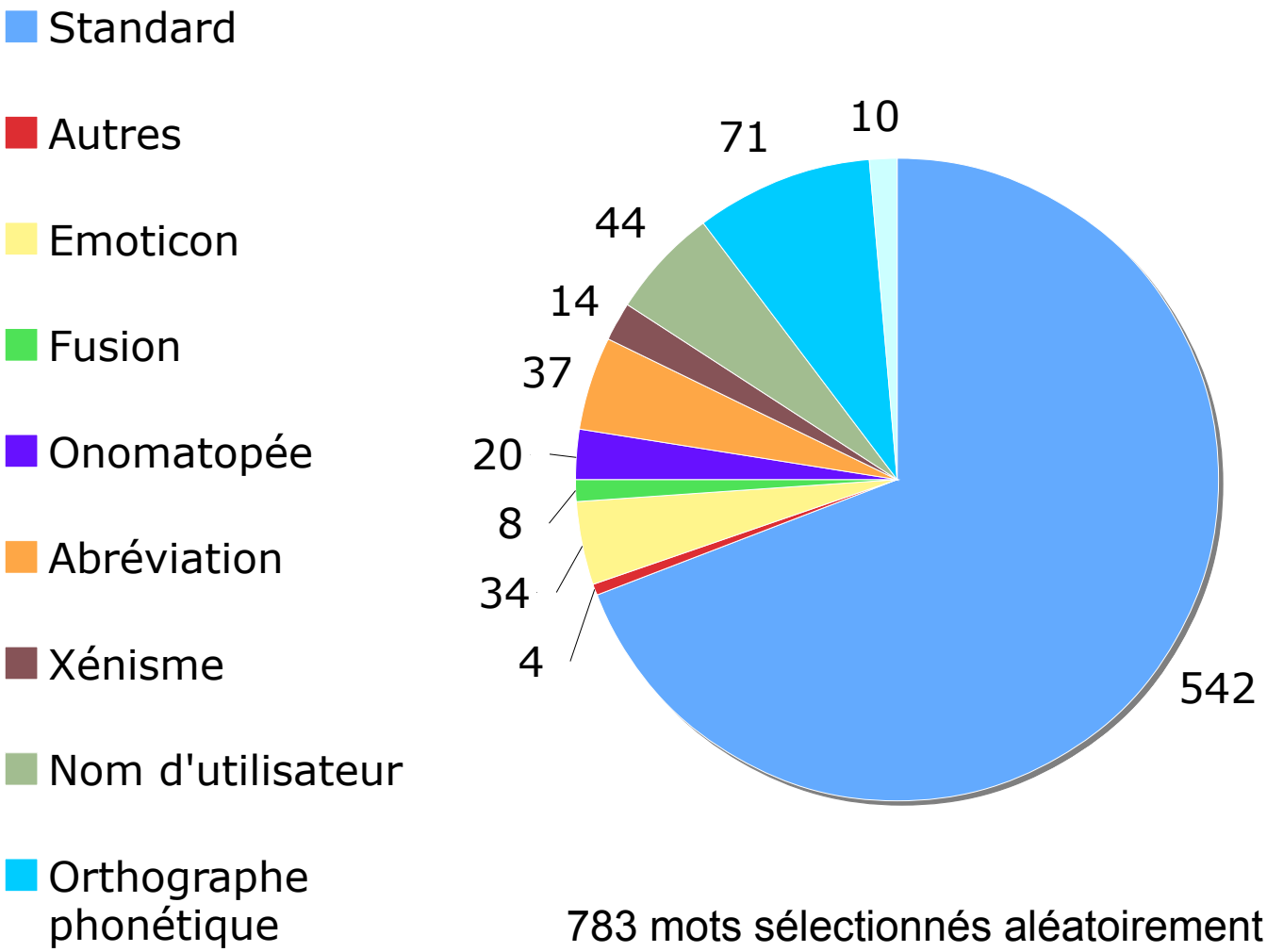
783 mots sélectionnés aléatoirement sur le canal #18-25ans

Abréviations

37 mots / 783

- *lol*
- *dsl*
- *tjrs*
- *‘lut*
- *bn*
- *++*

Aperçu du français tchaté



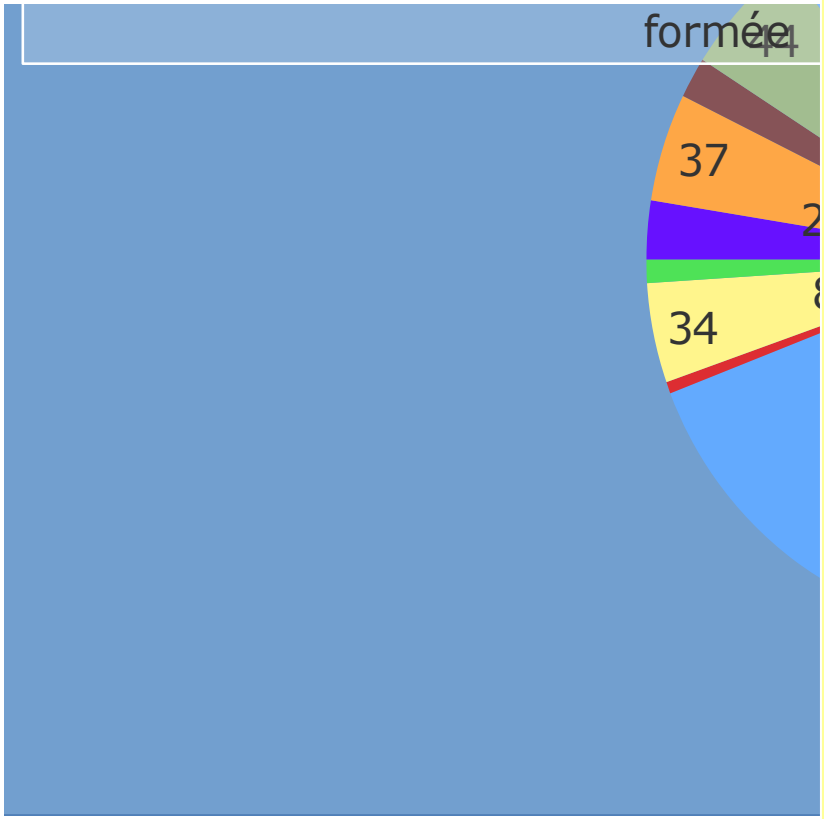
783 mots sélectionnés aléatoirement sur le canal #18-25ans

Emoticons

34 mots / 783

- *;)*
- *:)*
- *=))*
- *:-**
- *:p*

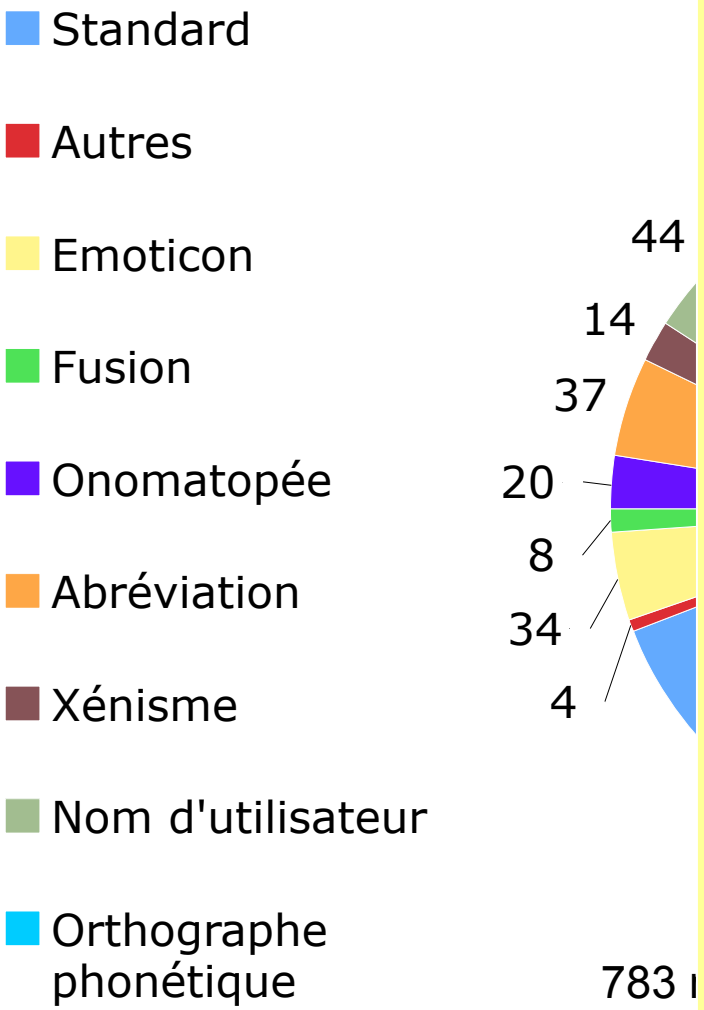
Aperçu du français tchaté



Corrections

A> *B* préfère sauce tarat
A> tartare*
B> du tou
A> sisi
A> a point
C> poivre rulez
D> sauce bernaise c la meilleure :)
A> frite moelleuse
C> nan a point sauce poivre
C> ça rox
A> lol
C> moelleuse mais croustillante stp

Aperçu du français tchaté

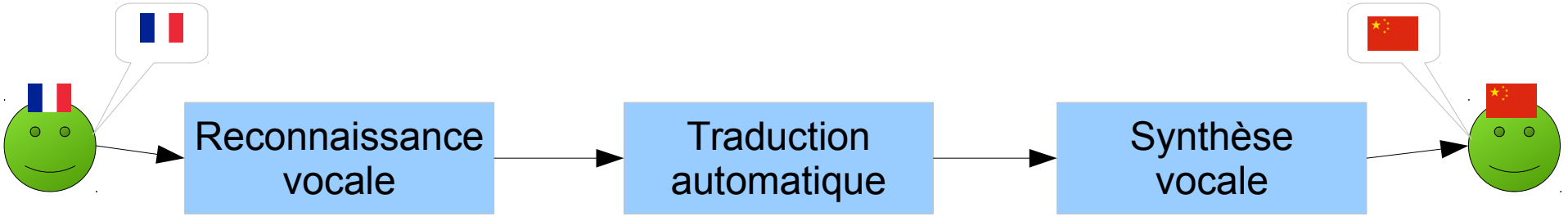


Corrections

A> *B* préfère sauce tarat
A> tartare*
B> du tou
A> sisi
A> a point
C> poivre rulez
D> sauce bernaise c la meilleure :)
A> frite moelleuse
C> nan a point sauce poivre
C> ça rox
A> lol
C> moelleuse mais croustillante stp

Intéressant à étudier
Possible à traiter

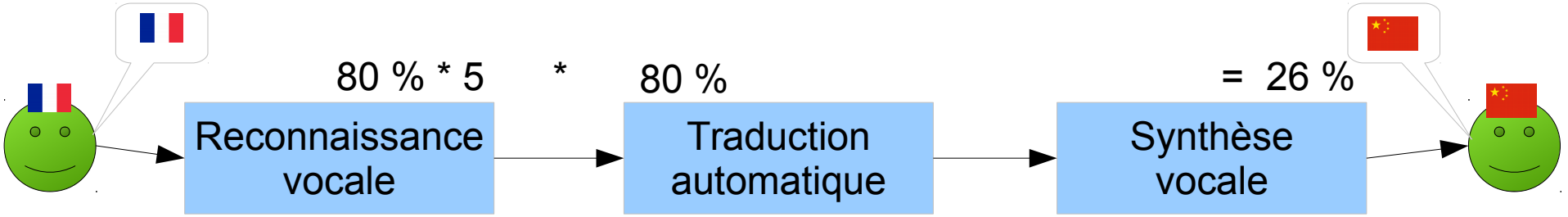
Industrial PhD (CIFRE) : spoken chat translation



- Nombreux projets scientifiques sur le sujet
  - C-STAR (1989-1993)
  - Verbmobil (1993-2000)
  - C-STAR II (1996-1999)
  - Nespole ! (2002-2004)

- Commercialisation ?
  - Prosodie (opérateur télécom grands comptes)

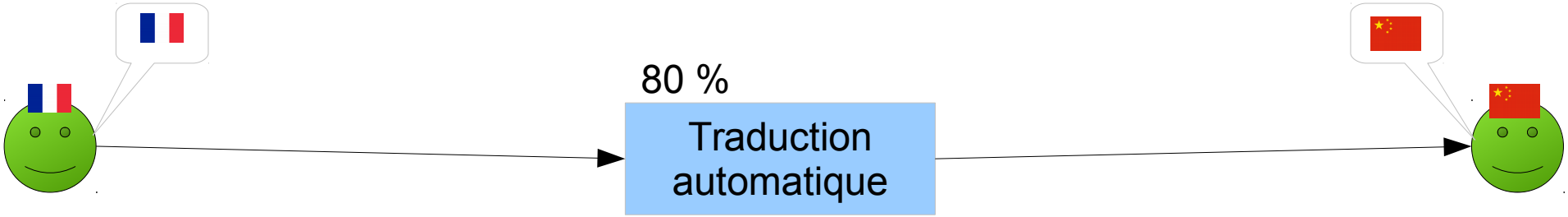
Industrial PhD (CIFRE) : spoken chat translation



- Nombreux projets scientifiques sur le sujet
  - C-STAR (1989-1993)
  - Verbmobil (1993-2000)
  - C-STAR II (1996-1999)
  - Nespole ! (2002-2004)

- Commercialisation ?
  - Prosodie (opérateur télécom grands comptes)

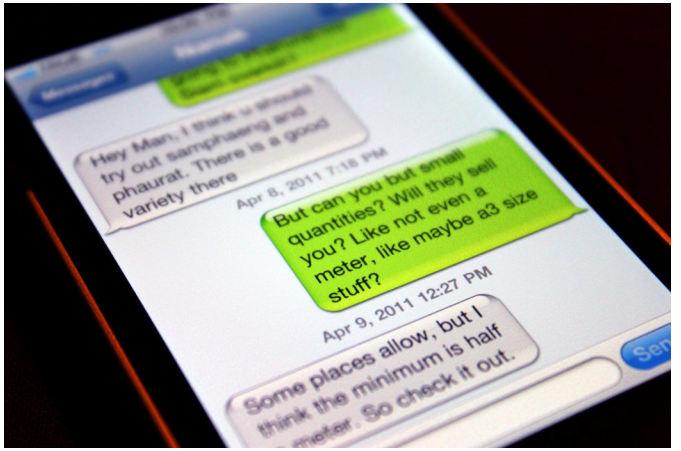
Industrial PhD (CIFRE) : ~~spoken~~ chat translation



- Nombreux projets scientifiques sur le sujet
  - C-STAR (1989-1993)
  - Verbmobil (1993-2000)
  - C-STAR II (1996-1999)
  - Nespole ! (2002-2004)

- Commercialisation ?
  - Prosodie (opérateur télécom grands comptes)

A second dimension ?

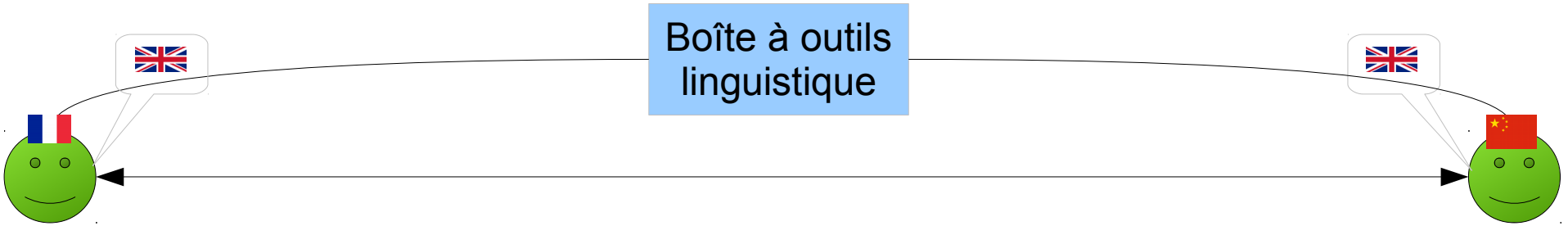
	Speech	Text
Conversation : Spontaneous Interactive		
Discourse : Planned One-sided		

What do people actually need ?

Travail sur corpus

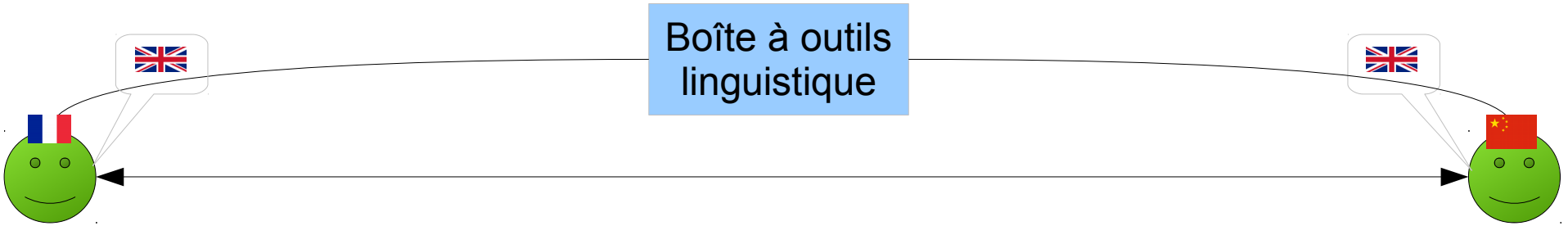
- Difficultés du dialogue en langue étrangère
- Stratégies de résolution des problèmes
  - Coopération, co-construction du sens
  - Métadialogue
- Importance du *feed-back*
  - Dans toutes les situations de dialogue

Industrial PhD (CIFRE) : assisted chat



- Boîte à outils
  - Traduction automatique
  - Dictionnaire intégré, tenant compte :
    - du contexte paradigmatique (mots ayant plusieurs définitions)
    - du thème (vocabulaire technique)
  - Dialogue métalinguistique
    - Coannotation des messages
    - Coédition des messages

Industrial PhD (CIFRE) : assisted chat



- Boîte à outils
  - Traduction automatique
  - Dictionnaire intégré, tenant compte :
    - du contexte paradigmatique (mots ayant plusieurs définitions)
    - du thème (vocabulaire technique)

Soutenance en 2009
- des spécifs, une démo, mais…
- pas d'XP
- pas de traitement (POS-tagging / lemmatisation) du tchat

Background

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017

Doctorat : aide au dialogue en langue 2nde — Informatique (Grenoble-1)

Thèse CIFRE (LIG-GETALP + Prosodie SA)

LIDILEM — Université Grenoble Alpes

Scientext Production/outillage de corpus arborés (LIDILEM) ScienQuest

OMNIA Annotation sémantique interlingue

(LIG-GETALP)

Traouiero Opérationnalisation, traduction de sites Web

ATER Dicorpus

Traitement du français classique (XVI^{e}-XVIII^{e} siècles) (ICAR) **Presto** CNRS ASLAN

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017

Projet ANR Scientext (2008-2009)

LIDILEM (EA 609, Grenoble)

Étude du positionnement et du raisonnement dans l'écrit scientifique

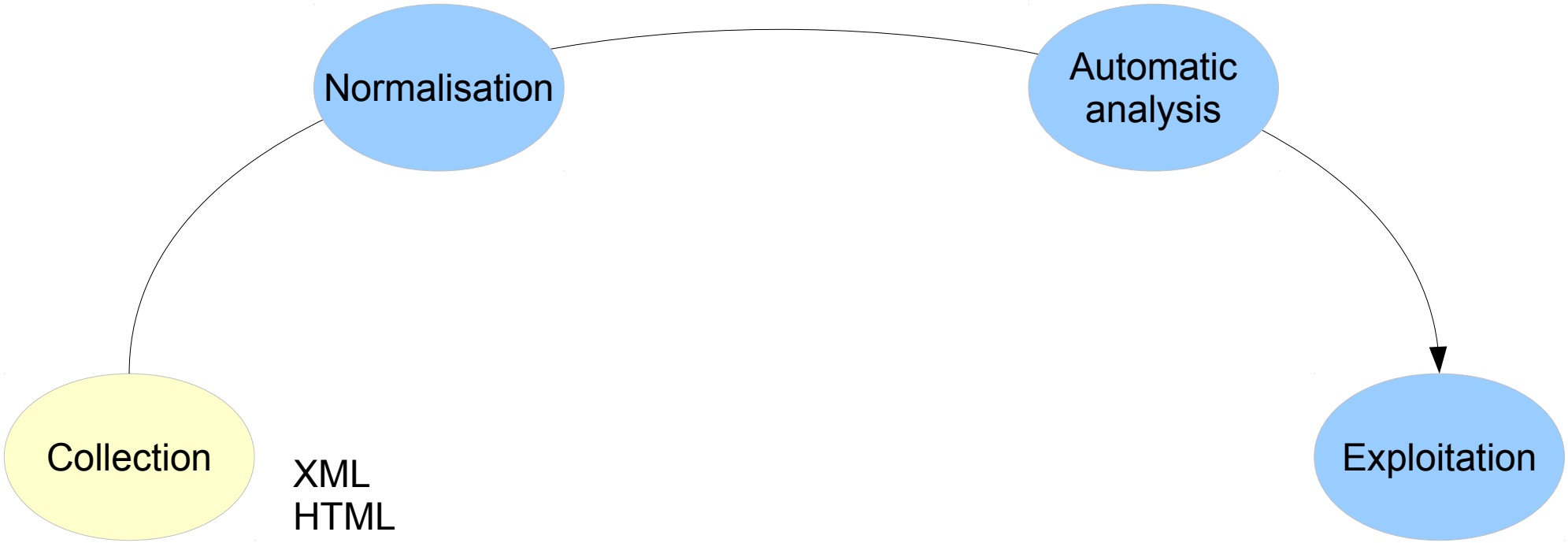
- Phraséologie
- Marques énonciatives
- Marques syntaxiques
- … liées à la causalité

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)

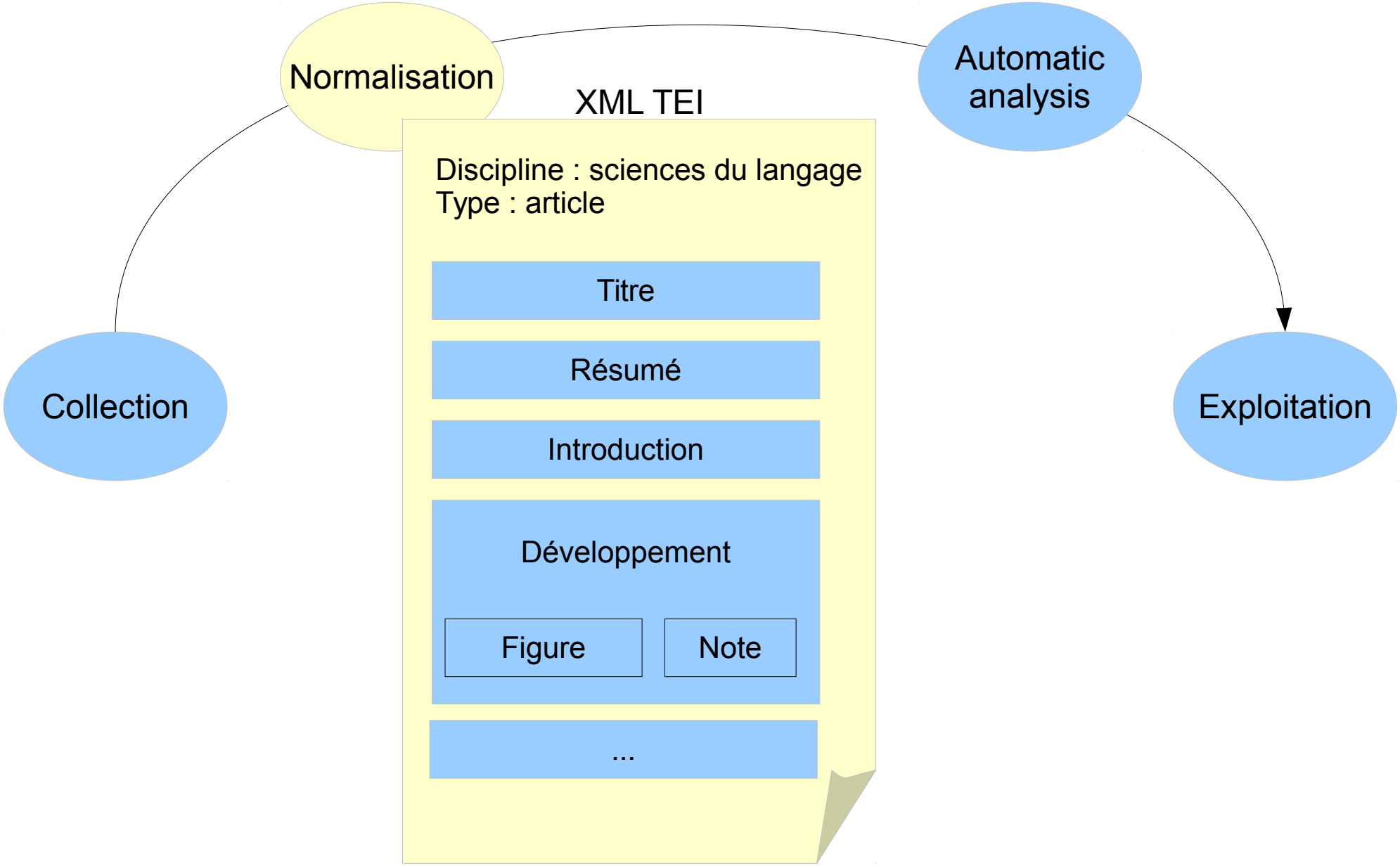
Projet ANR Scientext (2008-2009)

LIDILEM (EA 609, Grenoble)



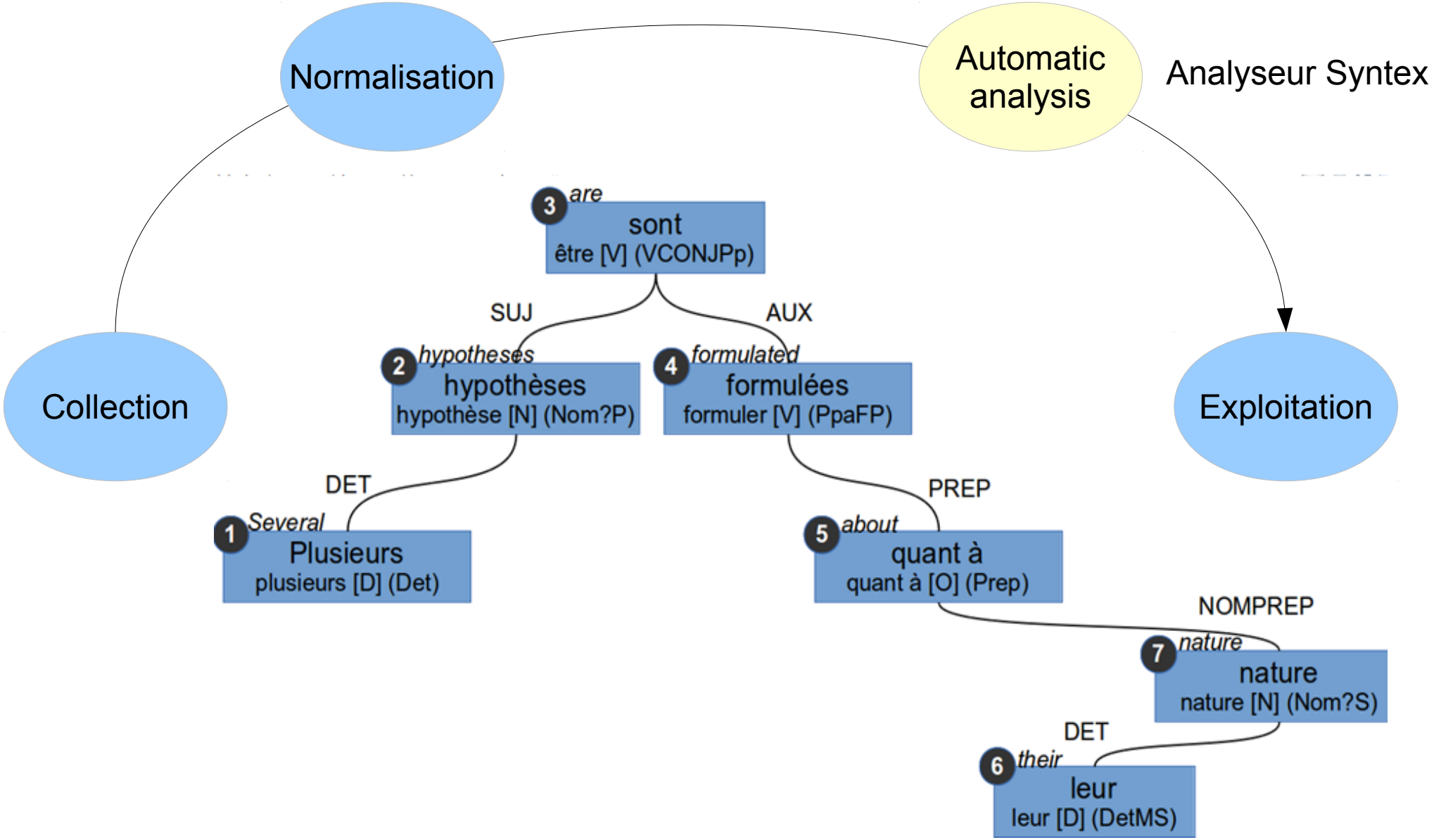
Projet ANR Scientext (2008-2009)

LIDILEM (EA 609, Grenoble)



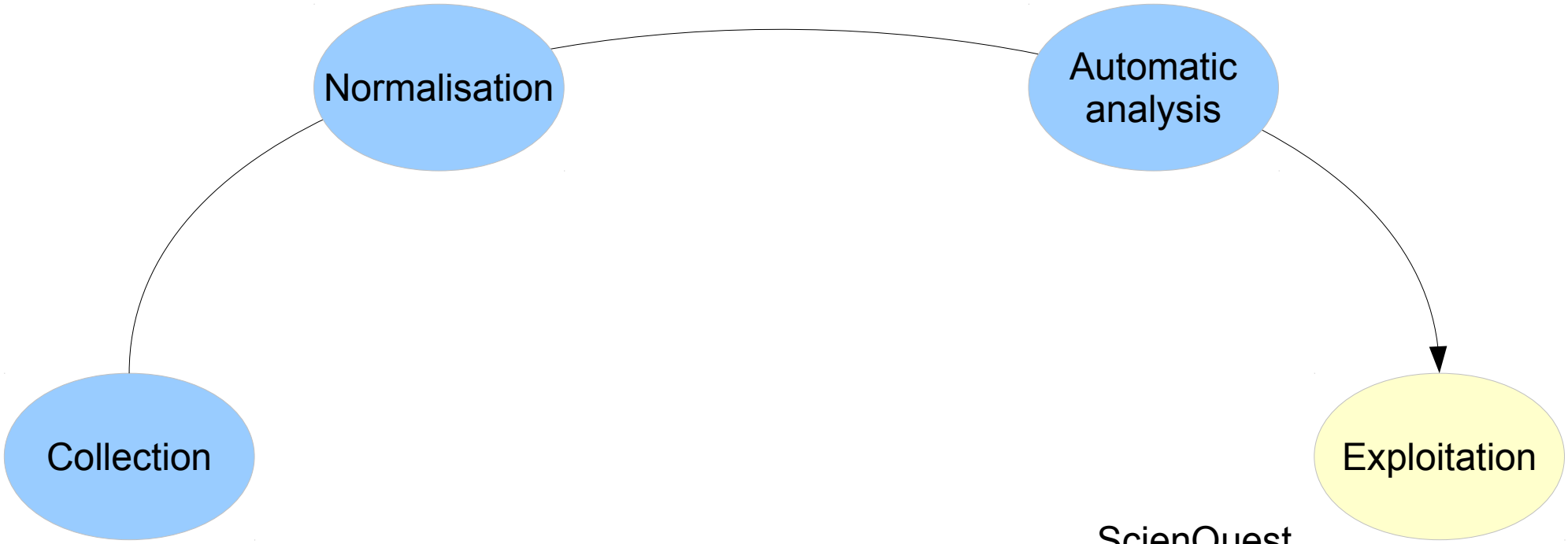
Projet ANR Scientext (2008-2009)

LIDILEM (EA 609, Grenoble)



Projet ANR Scientext (2008-2009)

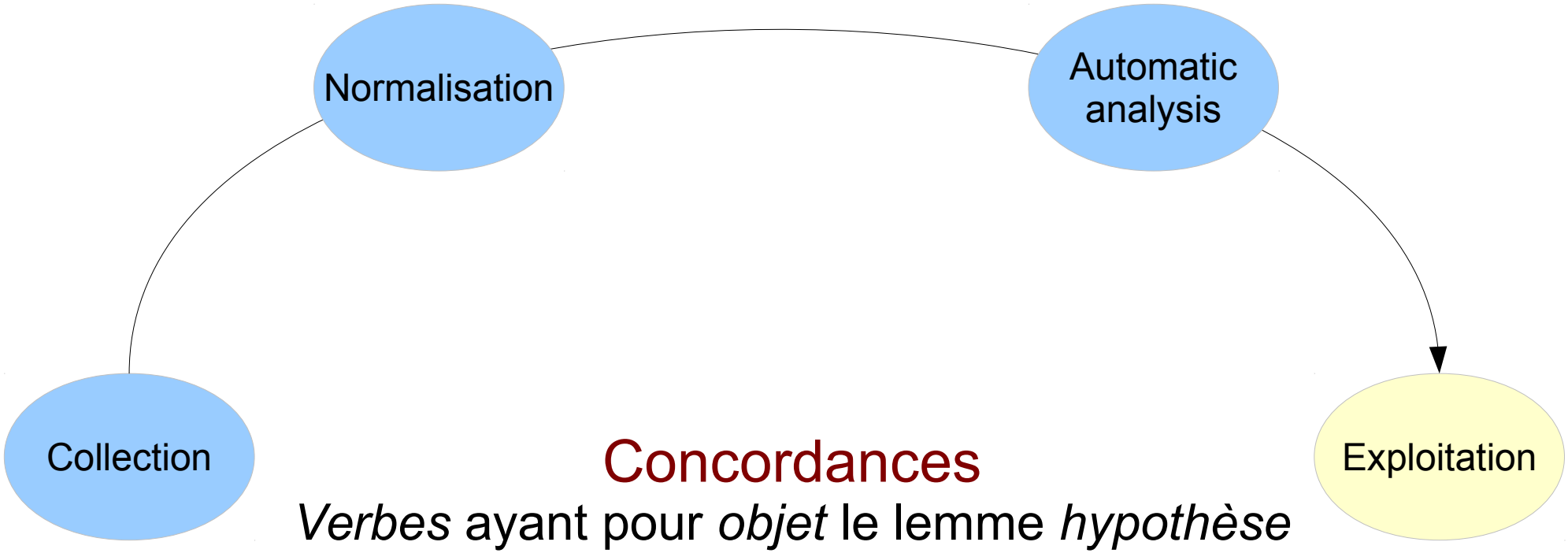
LIDILEM (EA 609, Grenoble)



ScienQuest
http://corpora.aiakide.net

Projet ANR Scientext (2008-2009)

LIDILEM (EA 609, Grenoble)



Concordances

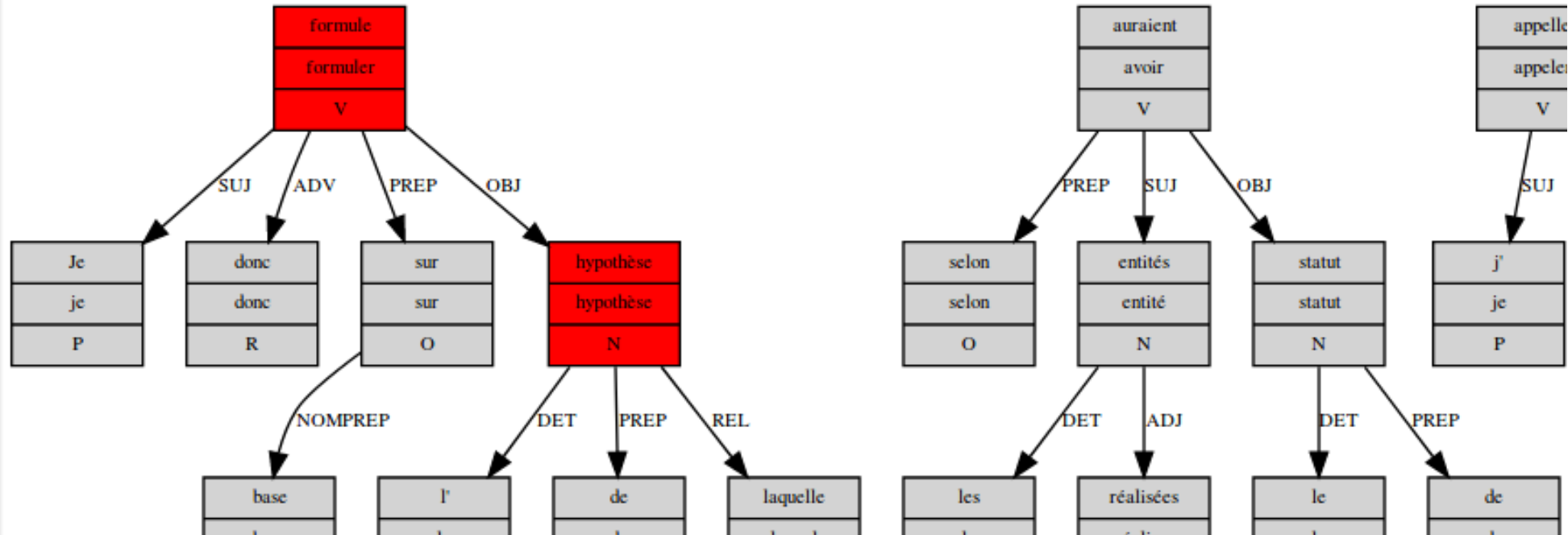
Verbes ayant pour *objet* le lemme *hypothèse*

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

Projet ANR Scientext (2008-2009)

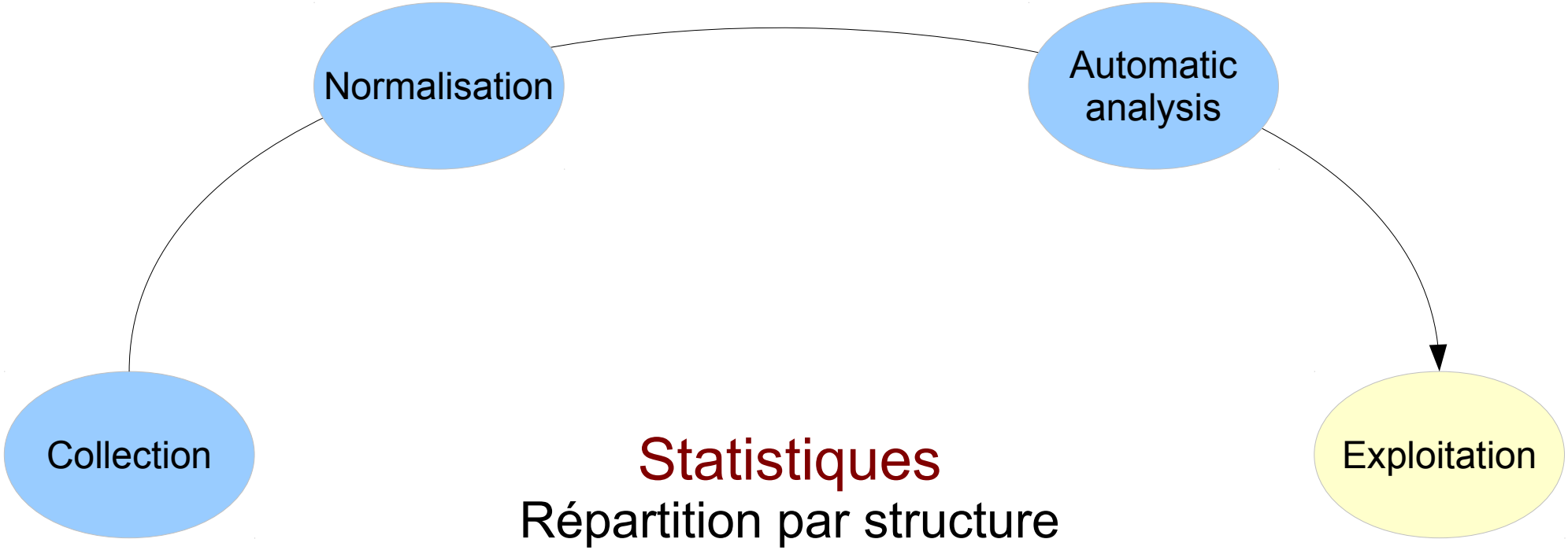
LIDILEM (EA 609, Grenoble)

Contexte gauche	Occurrences	Contexte droit	
marche de décomposition des savoirs et de création de situations	implique une hypothèse	forte : c' est que l' on peut prendre ,	#1 - Dév
iste une réponse simple en terme de commande si l' on	fait l' hypothèse	que la perturbation (couple d' adhérence) est constante	#2 -
sujets selon les types 3 et 4 , ce qui	exclut l' hypothèse	d' une détermination absolue du choix de formulation par les	#5 -
queurs visuels : ponctuation , typographie , disposition . Nous	faisons par ailleurs l' hypothèse	que ces marqueurs risquent de varier en fonction de paramètres liés au	#5 -
, par exemple la commande Distance On retrouve avec cette	analyse l' hypothèse	issue du modèle de représentation de l' architecture textuelle ,	#5 -
plie 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	#5 -
, à paraître : 11 - 12) . Je	formule donc sur cette base l' hypothèse	de travail selon laquelle les entités réalisées par ce que	#5 -



Projet ANR Scientext (2008-2009)

LIDILEM (EA 609, Grenoble)



Statistiques

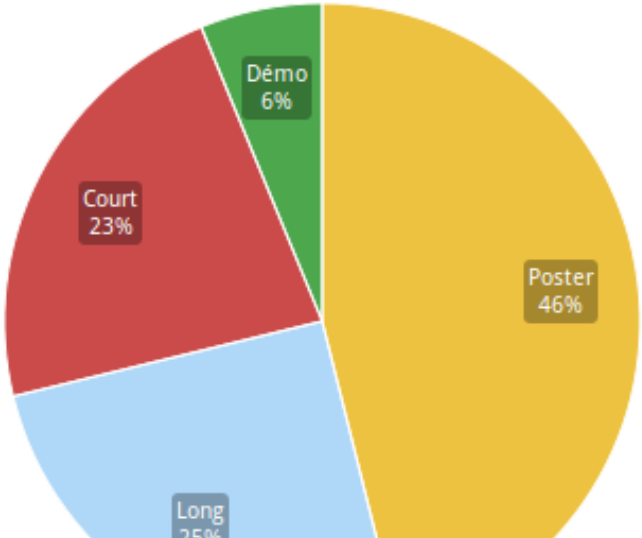
Répartition par structure

Afficher

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

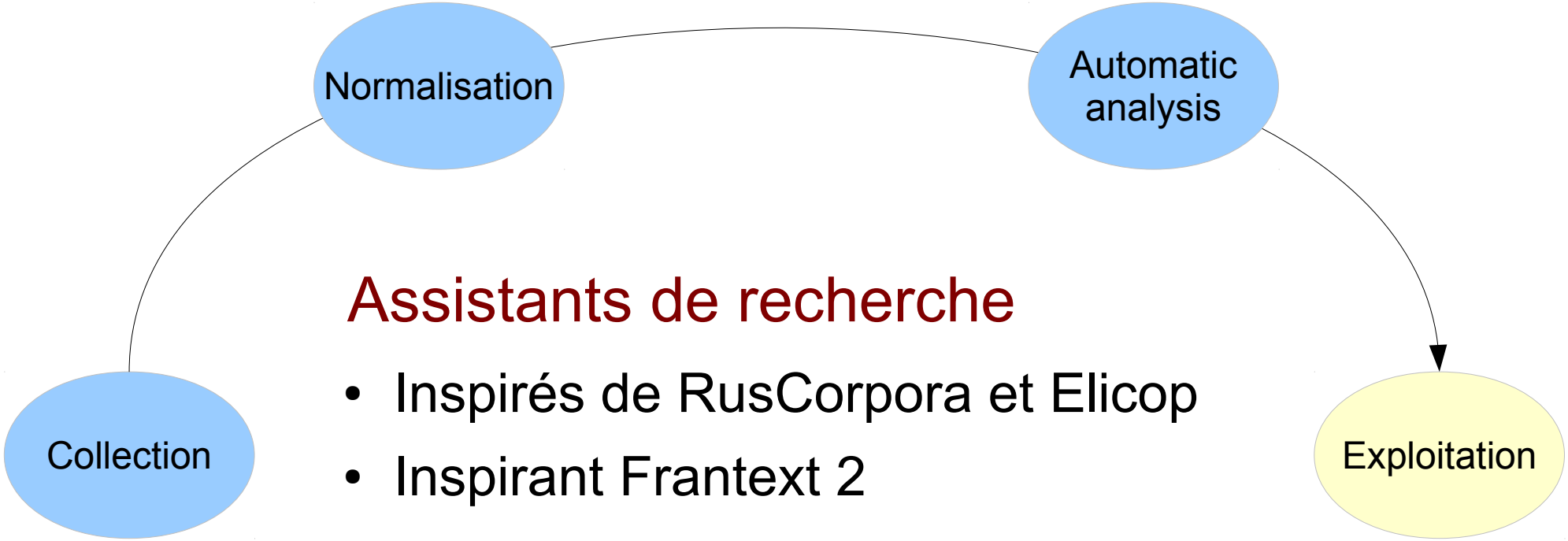
Type

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



Projet ANR Scientext (2008-2009)

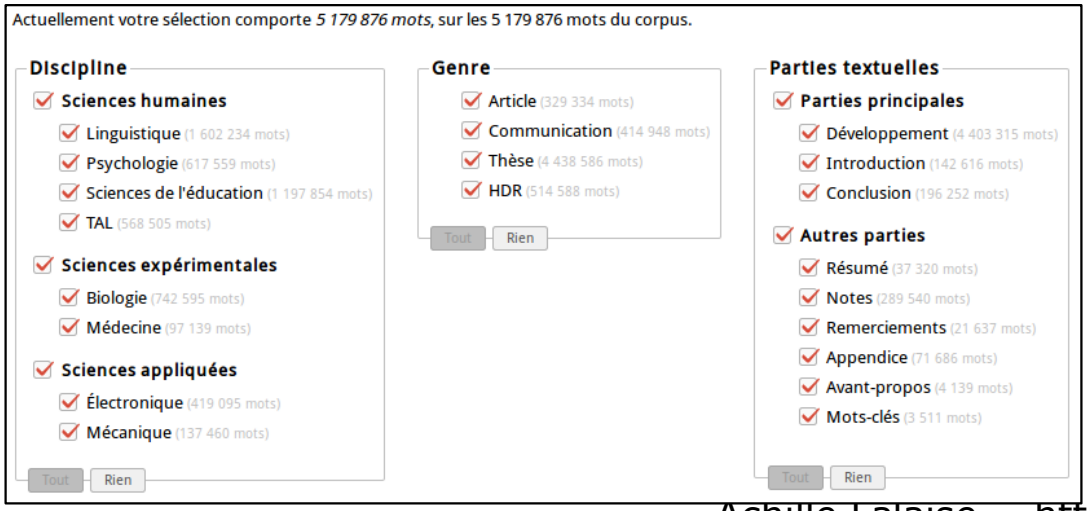
LIDILEM (EA 609, Grenoble)



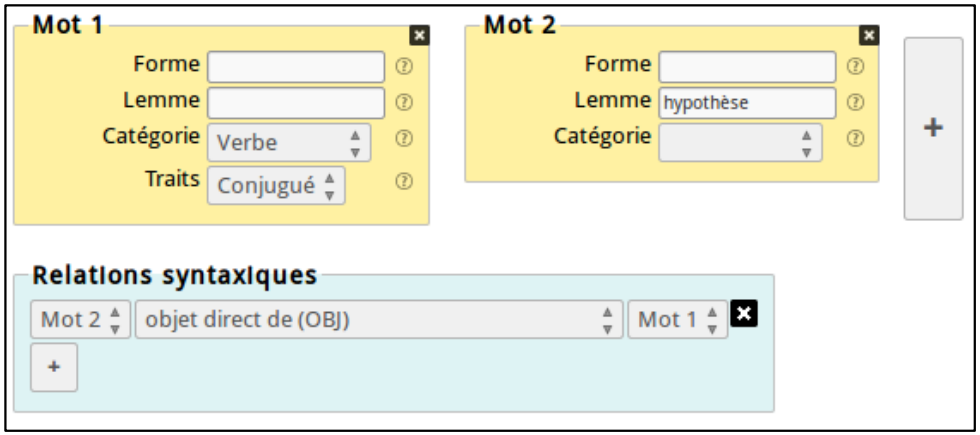
Assistants de recherche

- Inspirés de RusCorpora et Elicop
- Inspirant Frantext 2

Assistant de sélection structurale



Assistant de composition de requête



ScienQuest (2010-2018) : évolution

Évaluation avec des utilisateurs

- Formations : Paris (2014), Grenoble (2017), Paris (2017)...
- Démos : Toth 2010, TALN 2015, Valence 2016 et 2017, Grenoble 2018…
- TD : 2009 (M2 IdL Grenoble), 2014 (M1 IdL Grenoble), 2016 (doctorants SdL CUSO), 2017-2018 (L3 SdL Lyon)
- Retours d'utilisateurs

Évolution de l'outil

- Amélioration de l'ergonomie
- Ajout de fonctionnalités

  cf. *changelog* sur le site de ScienQuest
- Ajout de corpus

ScienQuest : stats d'utilisation 2016-2017

https://corpora.aiakide.net

~70 recherches / jour

- ~5 sessions / jour
  - Pics d'utilisation > 25 session simultanées
- ~14 recherches / session

À venir

- Stats ++
  - Mais comment rester simple et pédagogique ?
- Annotations/recherches portant au-delà de la syntaxe

Pays (Régions)	Sessions
France (ARA, IdF, Occitanie, Grand Est...)	50 %
Suisse (Neuchâtel, Vaud, Zurich...)	10 %
Tunisie (Tunis...)	9 %
Allemagne (Saxe, Saxe-Anhalt, Thuringe...)	4 %
Luxembourg	2 %
Espagne (Andalousie...)	2 %
Algérie (Batna, Alger...)	2 %
Canada (Québec, Ontario...)	2 %

ScienQuest : private Corpora

Année	Financement	Type	Langue	Nb mots	Analyse
2011	ANR ÉMOLEX	Littérature, journaux	Allemand	301M	Connexor
			Anglais	200M	XIP
			Espagnol	286M	Connexor
			Français	230M	Connexor
			Russe	554k	DeSR
2013	ANR Termith	Sciences humaines	Français	5M	XIP
2016	ANR Presto	Littérature, lettres, essais, traités	Français (XVI^{e}–XXI^{e} s)	35M	Presto
		Le Monde (échantillon, 1945-2015)	Français	16M	Presto
		Offices de tourisme (722 sites Web)	Français	4M	Presto
		Trip Advisor (échantillon)	Français	4M	Presto
		Wikitravel	Français	5M	Presto
2017	Labex TransferS	Articles et communiqués de presse, textes juridiques, sites Web	En/Es/Fr		CoreNLP
2017	CORLI	Articles sc. et transcriptions de conférences	En/Fr		Talismane

Background

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017

Doctorat : aide au dialogue en langue 2nde — Informatique (Grenoble-1)

Thèse CIFRE (LIG-GETALP + Prosodie SA)

Scientext Production/outillage de corpus arborés (LIDILEM) ScienQuest

(LIG-GETALP)

OMNIA Annotation sémantique interlingue

Traouiero Opérationnalisation, traduction de sites Web

ATER — Dicorpus

Traitement du français classique (XVI^{e}-XVIII^{e} siècles) (ICAR) **Presto** CNRS ASLAN

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017

OMNIA : annotation sémantique pour la RI

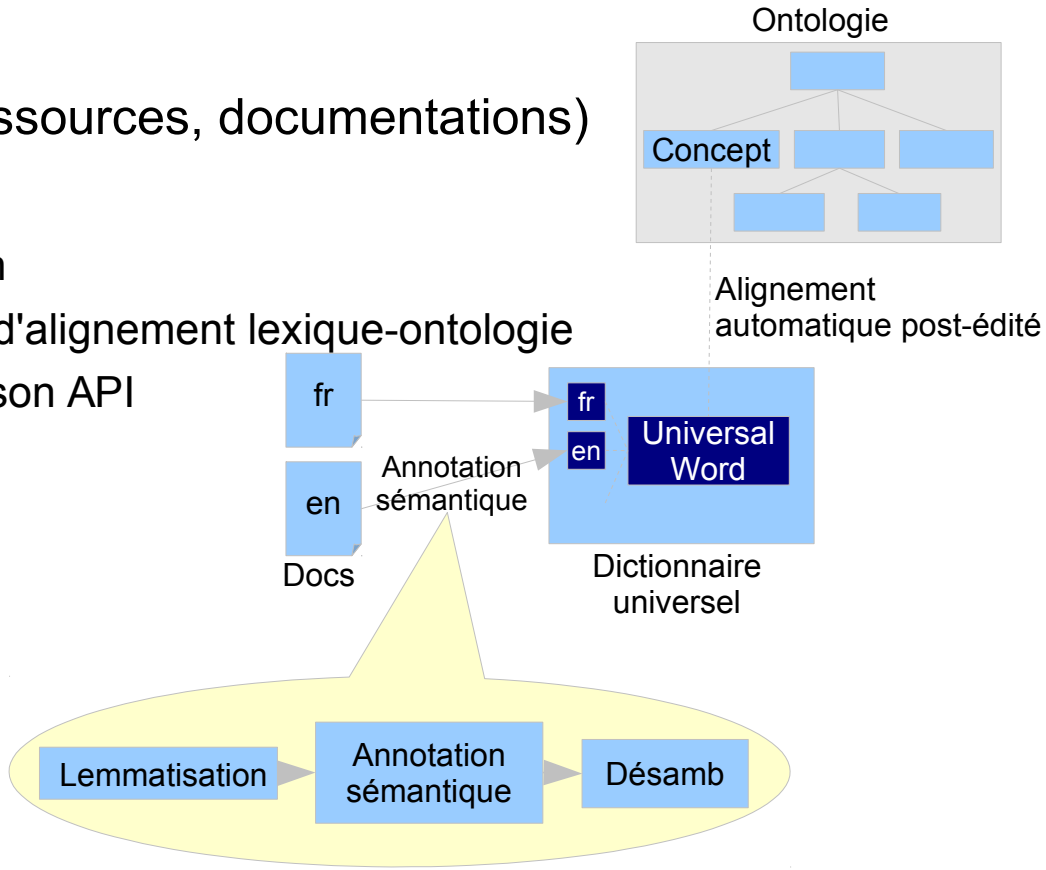
ANR : XEROX (Grenoble), LIG-GETALP (Grenoble-1), LIRIS (EC Lyon)

Situation

- Tâche : indexation de textes indépendamment de la langue
- Difficulté à intégrer du lexique (*objet linguistique*) dans des ontologies (*concept*)
- Approche par dictionnaire, outils et ressources déjà existants dans l'équipe

Méthode

- Inventaire des composants (outils, ressources, documentations)
- Création
  - du composant manquant : lemmatisation
  - d'une ontologie de test et du processus d'alignement lexique-ontologie
  - du format d'échange (graphes-Q) et de son API
  - de la chaîne de traitement

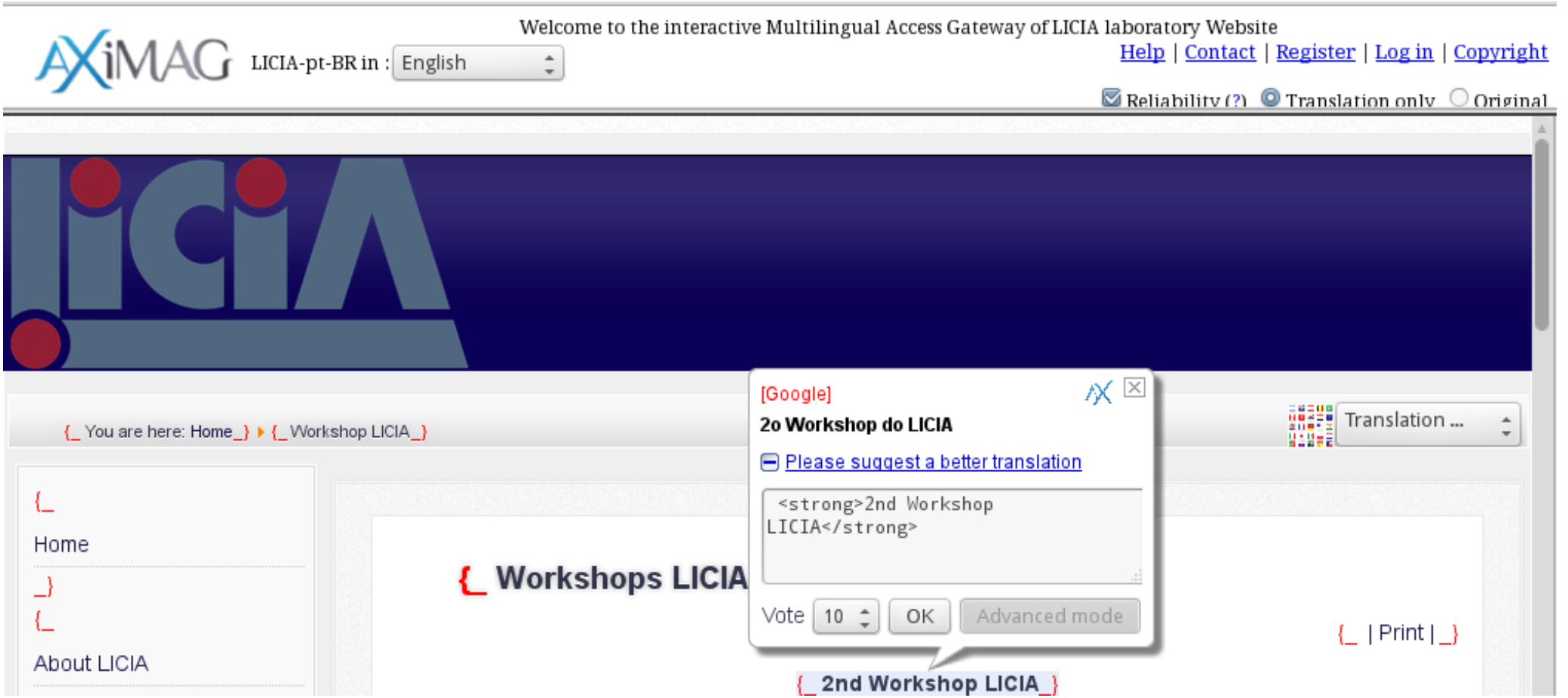


Logiciels et ressources : Linux, Apache, Tomcat, MySQL, Eclipse, Subversion, DELA, DELAF, Wordnet, Pivax, Jibiki

Langages : Java, PHP, Perl, Systèmes-Q, UNL, RDF, OWL

Projet ANR Traouiero (2011-2012)

LIG-GETALP (UMR 5217, Grenoble), Floralis (Grenoble)



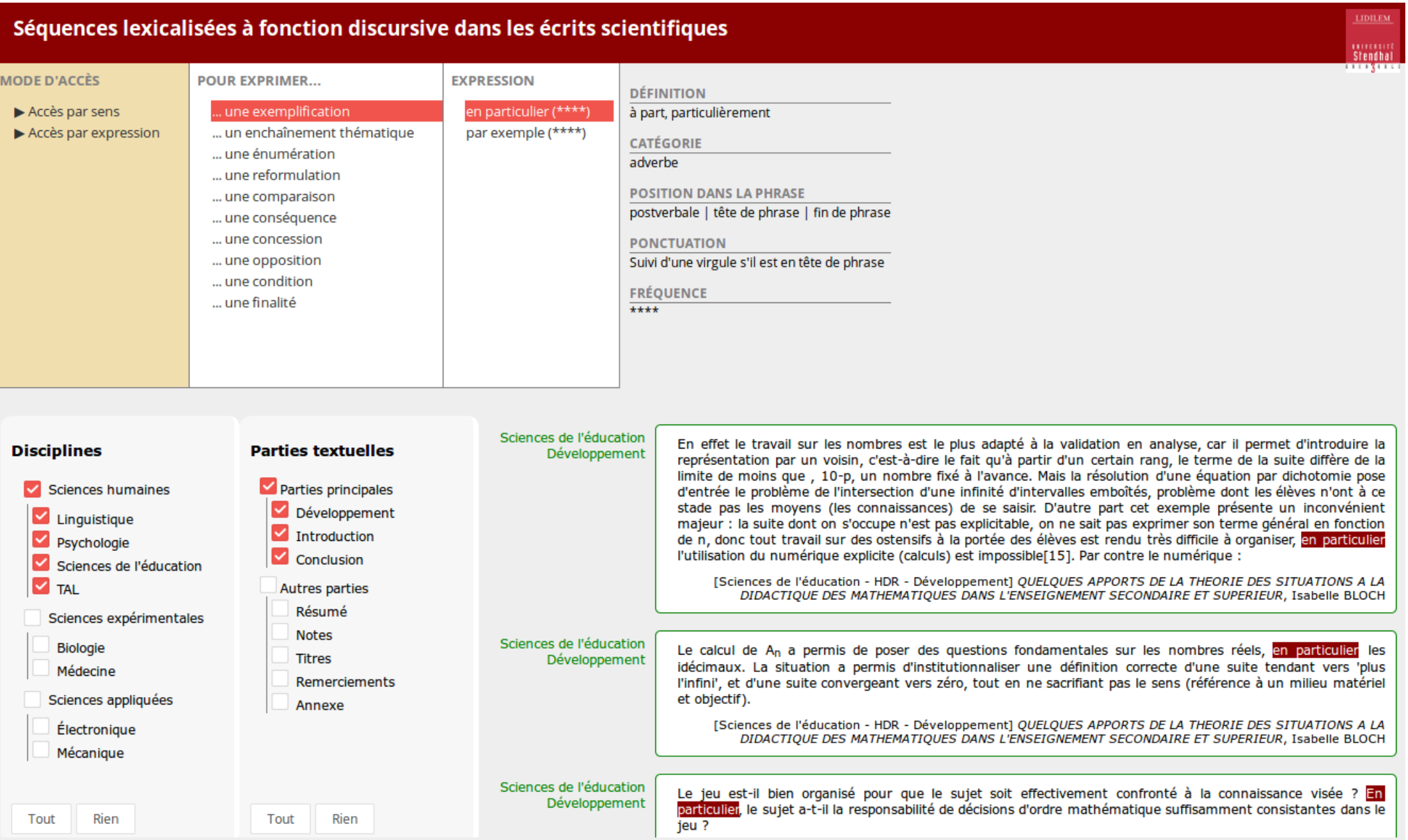
Dicorpus : base de données lexicales basée sur corpus

- Expressions polylexicales
  - absentes des dictionnaires
  - très fréquentes dans les textes scientifiques

En premier lieu, nous souhaitons défendre la thèse selon laquelle les expressions polylexicales répondent à des régularités. Nous remettons en question la thèse anomaliste et proposons jusqu'à un certain point un point de vue analogiste.

- Utilisation de Dicorpus pour le FOU (français sur objectifs universitaires)
- Travaux existants sur l'utilisation de corpus pour l'enseignement des langues

Dicorpus : base de données lexicales basée sur corpus



Dicorpus : base de données lexicales basée sur corpus

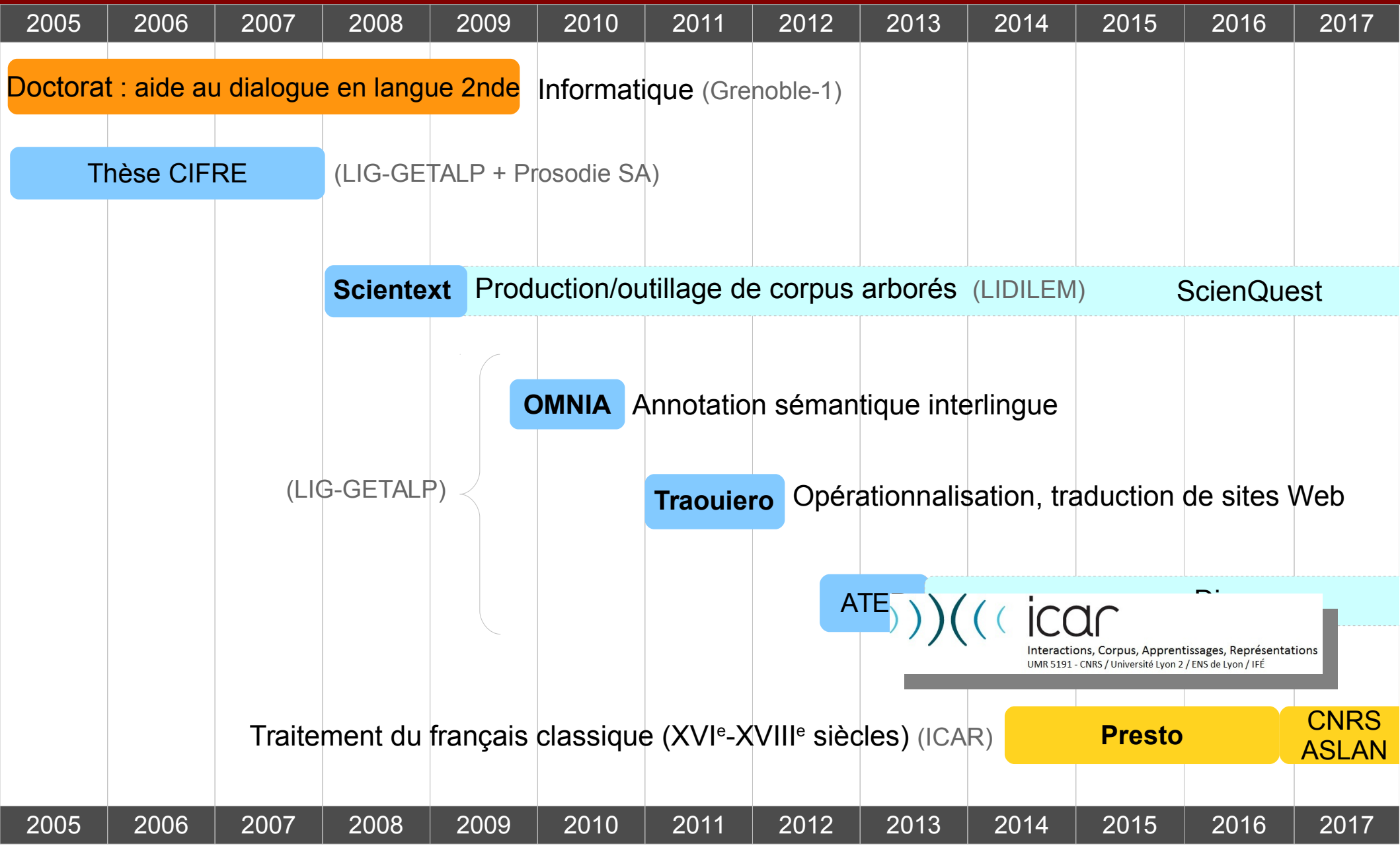
Travail avec

- Hoai Tran (Grammatica, U. Artois)
- Agnès Tutin (LIDILEM, U. Grenoble-Alpes)
- Cristelle Cavalla (DILTEC, Paris-3)

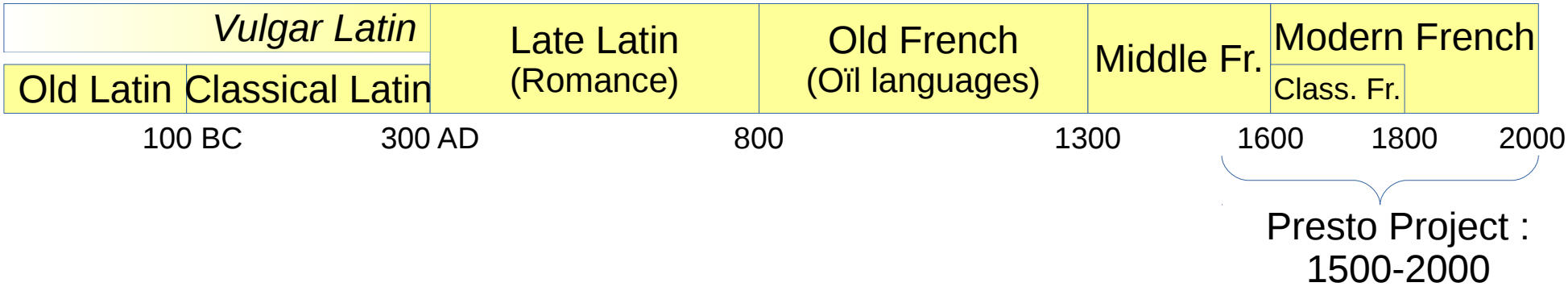
Plein de questions

- Ergonomie de la recherche
- Présentation des exemples
- Évaluation

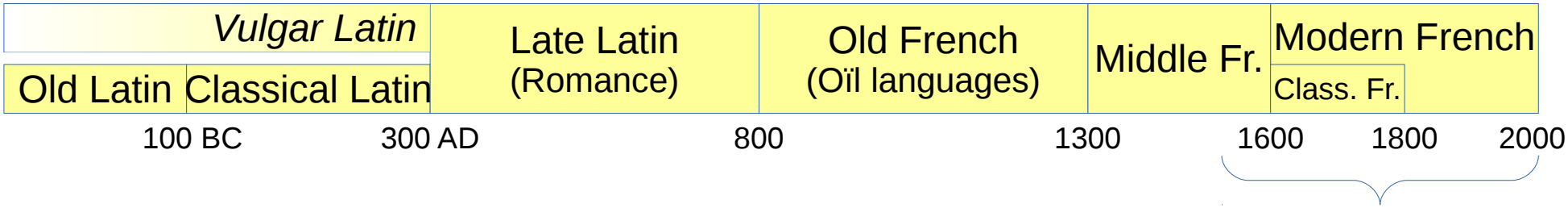
Background


2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Doctorat : aide au dialogue en langue 2nde
Informatique (Grenoble-1)
Thèse CIFRE
(LIG-GETALP + Prosodie SA)
Scientext
Production/outillage de corpus arborés (LIDILEM)
ScienQuest
OMNIA
Annotation sémantique interlingue
(LIG-GETALP)
Traouiero
Opérationnalisation, traduction de sites Web
ATE
icar
Interactions, Corpus, Apprentissages, Représentations
UMR 5191 - CNRS / Université Lyon 2 / ENS de Lyon / IFÉ
Traitement du français classique (XVIe-XVIIIe siècles) (ICAR)
Presto
CNRS
ASLAN
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

1500-1800 : the rise of French as a national, standardised language



1500-1800 : the rise of French as a national, standardised language



- 1440 : Printing Press
- 1532 : *Pantagruel*, François Rabelais
- 1539 : Ordinance of Villers-Cotterêts − French is now France's official language
- 1543 : *Traité des reliques*, Jean Calvin
- 1549 : *La Deffence et Illustration de la Langue Francoyse*, Joachim du Bellay
- 1550 : *Briefve collection de l'administration anatomique*, Ambroise Paré
- 1562 : *Gramere*, Pierre de la Ramée
- 1572-1592 : *Essais*, Montaigne
- 1634 : Académie française − state-sponsored standardisation of French langage
- 1751-1772 : *Encyclopédie*, Diderot & d'Alembert
- 1795 : École Normale de l'an III − standardisation of French education

XIIth Century French

La Chanson de Roland

Quant l'Emperere ad faite sa justice
E esclargiée est la sue granz ire,
En Bramimunde ad chrestientet mise.
Passet li jurz, la nuiz est aserie,
Culchet s'est li Reis en sa cambre voltice.
Seinz Gabriel de part Deu li vint dire :
« Carle, semun les oz de tun empire,
« Par force iras en la tere de Bire,
« Rei Vivien si succurras en Imphe,
« A la citet que païen unt asise.
« Li chrestien te recleiment e crient. »
Li Emperere n'i volsist aler mie :
« Deus ! dist li Reis, si penuse est ma vie ! »
Pluret des oilz, sa barbe blanche tiret...

XVIth Century French

SI LA NATURE (dont quelque Person- de grand'renommée non sans rayson a douté, si on la devoit appeller Mere, ou Maratre) eust donné aux Hommes un commun vouloir, & consentement, outre les innumerables commoditez, qui en feussent procedées, l'Inconstance humaine, n'eust eu besoing de se forger tant de manieres de parler. Laquéle diversité, & confusion, se peut à bon droict appeller la Tour de Babel.

Début de *La deffence, et illustration de la langue francoyse*, Joachim du Bellay, 1549.

XVIIIth Century French

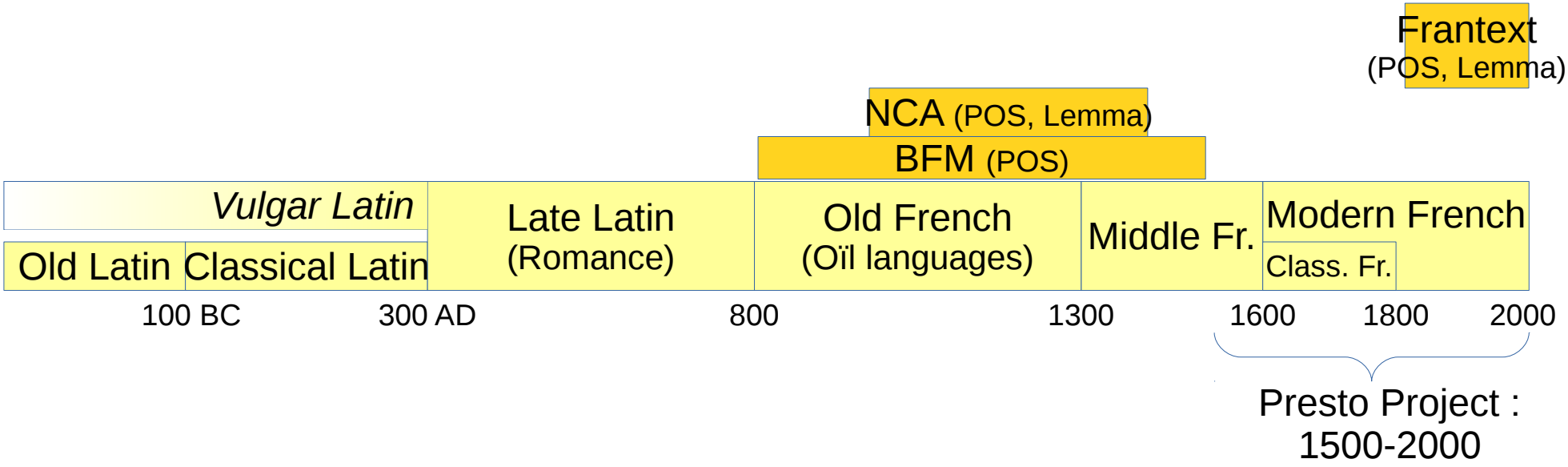
LANGAGE, s. m. (Arts. Raisonn. Philos. Metaphys.)

modus & usus loquendi, **maniere** dont les hommes se communiquent leurs pensées, par une suite de paroles, de gestes & d'expressions adaptées à leur génie, leurs **moeurs** & leurs climats.

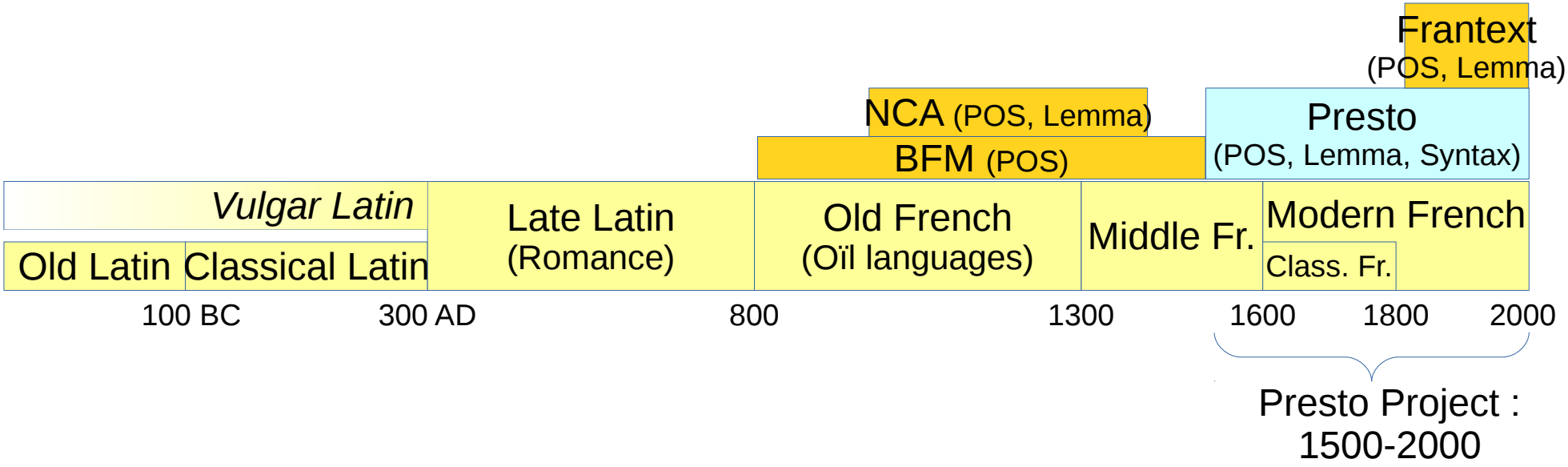
Dès que l'homme se sentit entraîné par goût, par besoin & par plaisir à l'union de ses semblables, il lui **étoit** nécessaire de développer son **ame** à un autre, & lui en communiquer les situations. Après avoir essayé plusieurs sortes d'expressions, il s'en tint à la plus naturelle, la plus utile & la plus étendue, celle de l'organe de la voix. Il **étoit aise** d'en faire usage en toute occasion, à chaque instant, & sans autre peine que celle de se donner des **mouvemens** de respiration, si doux à l'existence.

Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, Diderot & d'Alembert (dir.), 1751-1765

Tagged diachronic corpora for French



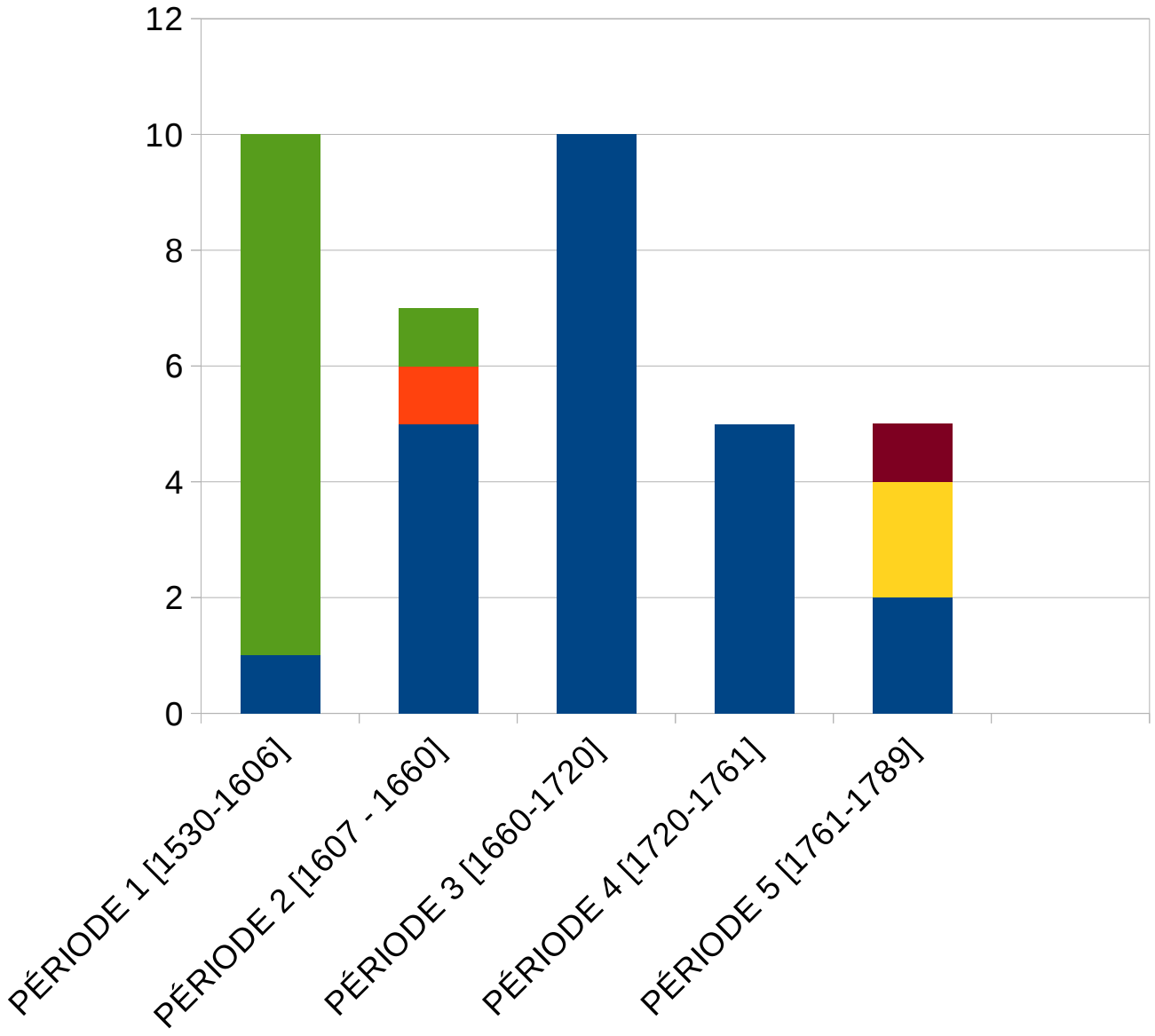
Tagged diachronic corpora for French



- Presto corpus
  - Litterature, essays, memoirs, etc.
  - 339 texts (of which 53 under CC-BY-NC licence)
  - 29M words

Tagged diachronic corpora for French

- P
  -
  -
  -



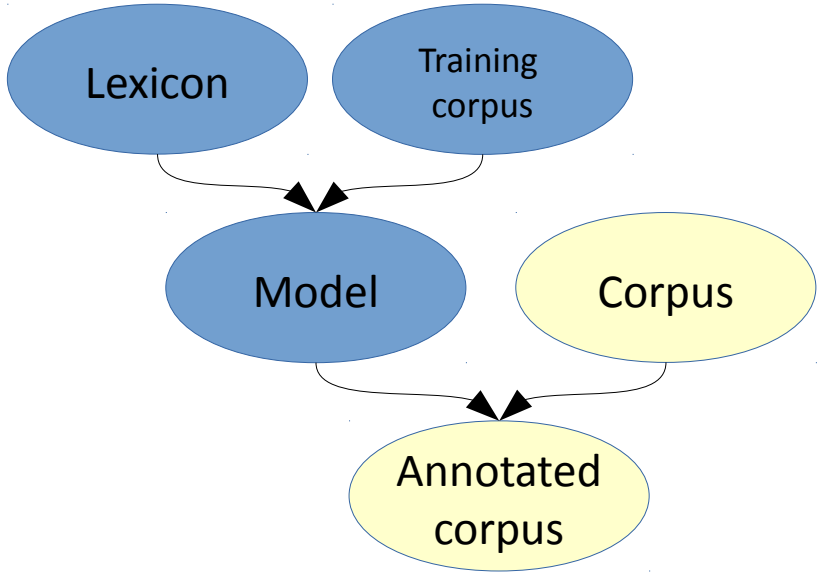
Old La

Frantext
S, Lemma)
Syntax)
French
0 2000
oject :
000

Jeu d'étiquettes (basé sur Multext/Eagles + Grace)

Étiquette niv 1	Étiquette niv 2	Flexion
Nom (N)	commun, propre	
Verbe (V)	être/avoir, autre	conjugué, infinitif
Adjectif (A)	général, possessif	
Pronom (P)	personnel, démonstratif, indéfini, possessif, interrogatif, relatif	
Déterminant (D)	article défini, démonstratif, article indéfini, article partitif, indéfini, relatif, interrogatif/exclamatif	
Participe-Adjectif-Gérondif (G)	part_présent/adjectif_verbal/gérondif, part_passé/adjectif_verbal	
Adverbe (R)	général, particule, interro-excalamatif	
Adposition (S)		
Conjonction (C)	coordination, subordination	
Numéral (M)	cardinal, ordinal	
Interjection (I)		
Résidu (X)	abréviation, mot étranger, symbole, préfixe, consonne intercalée	
Ponctuation (F)	forte, faible, autre	

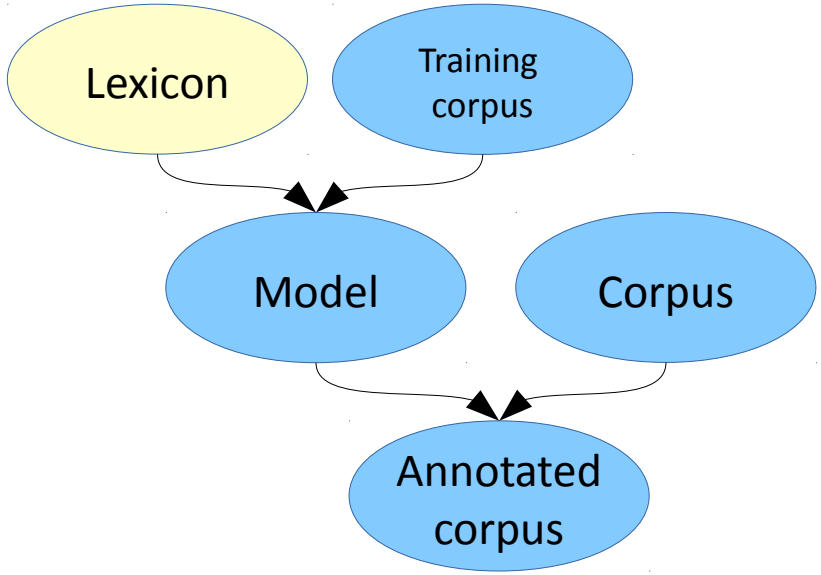
A classic workflow



Tokens, POS, lemmas

- Constraints :
  - *No feedback* between agents
  - *Tokens* are the minimal unit
    => tokenisation ambiguities are not allowed

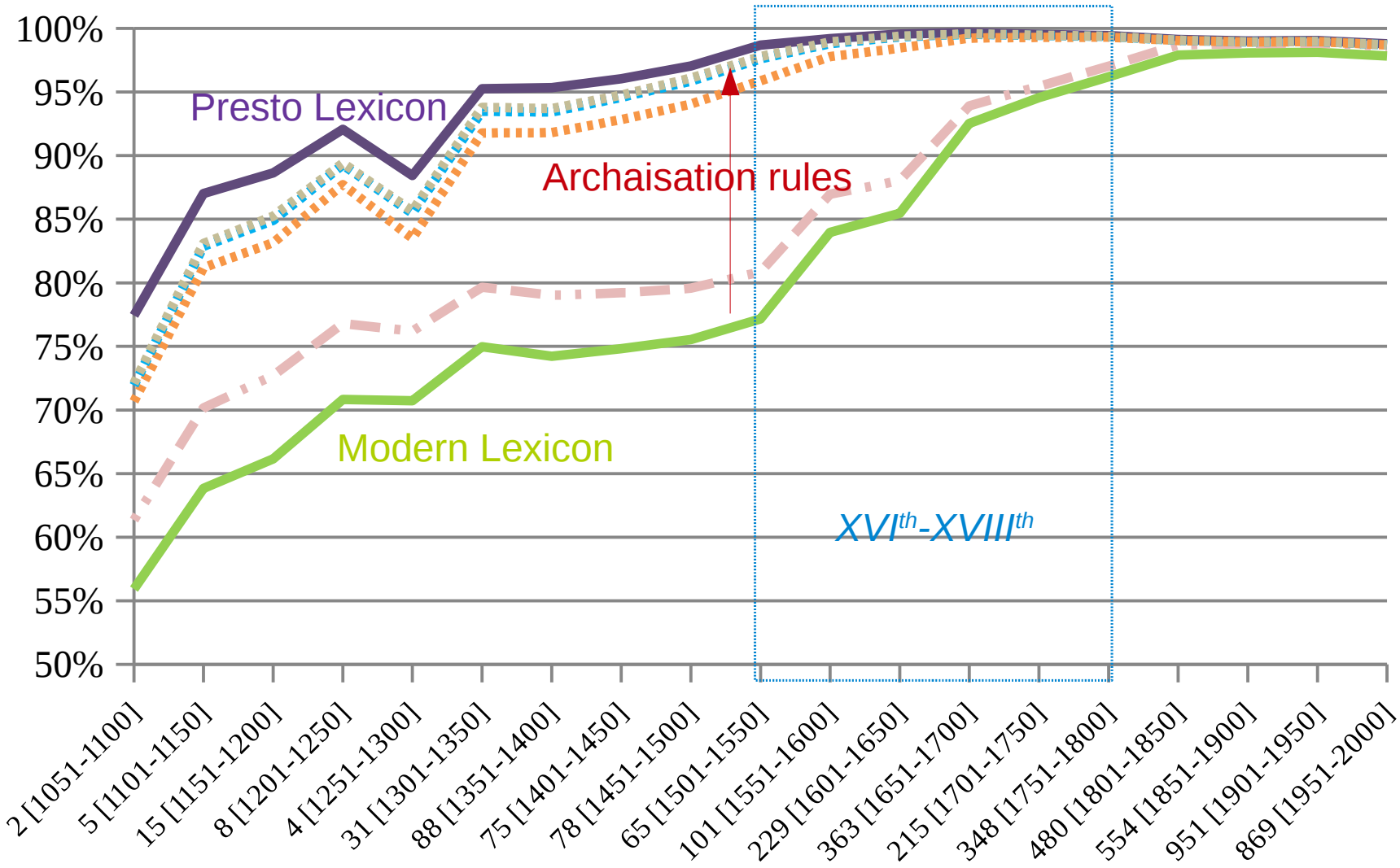
A classic workflow



Tokens, POS, lemmas

- Constraints :
  - *No feedback* between agents
  - *Tokens* are the minimal unit
    => tokenisation ambiguities are not allowed

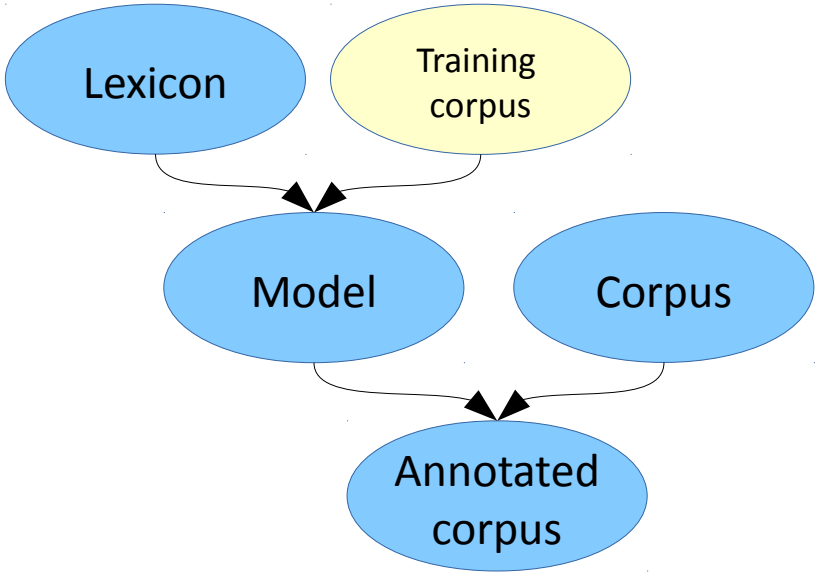
Lexical coverage



Couverture lexicale, mesurée sur le corpus *Frantext*, pour le lexique moderne (vert), les 3 itérations d'archaïsation (pointillés), et le lexique Presto final (violet).

Gilles Souvay,
ATILF, Nancy

A classic workflow

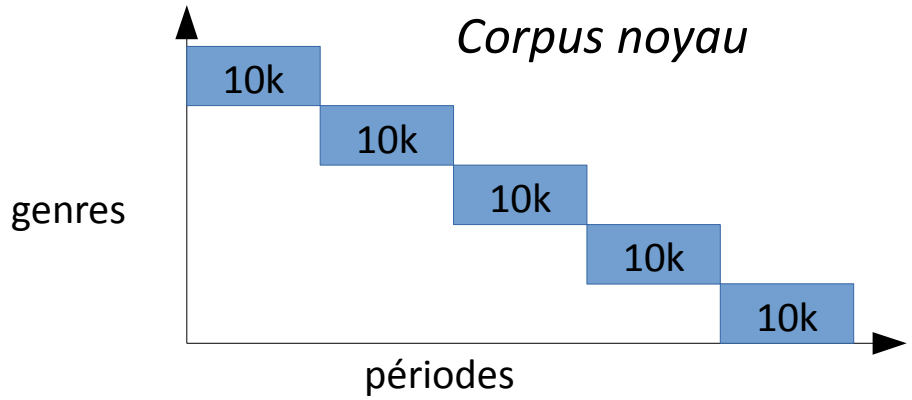


Tokens, POS, lemmas

Corpus d'apprentissage

Sélection de 5 textes

- 5 périodes
- 5 genres



Total : 62k tokens

Saulsaye (1547)

Lisandre et Caliste (1631)

Les Lettres de messire Roger de Rabutin, comte de Bussy (1681)

Essay sur l'histoire generale et sur les moeurs et sur l'esprit des nations (1756)

Le Paysan perverti ou les Dangers de la ville (1776)

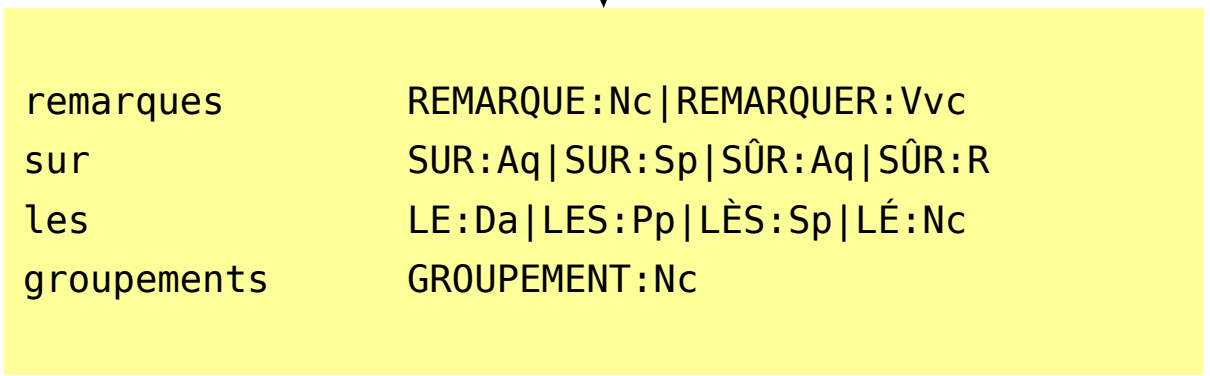
Projection lexicale

Lexique

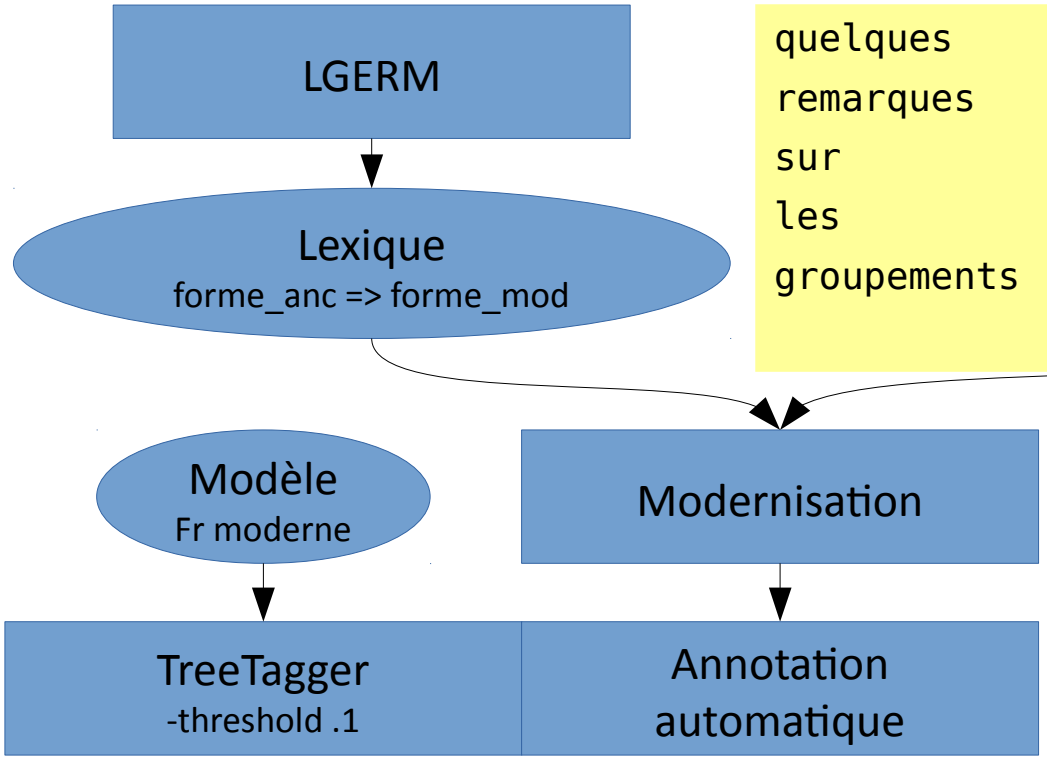
remarques → REMARQUE, Nc
remarques → REMARQUER, Vvc

Texte à anoter

Remarques sur les groupements

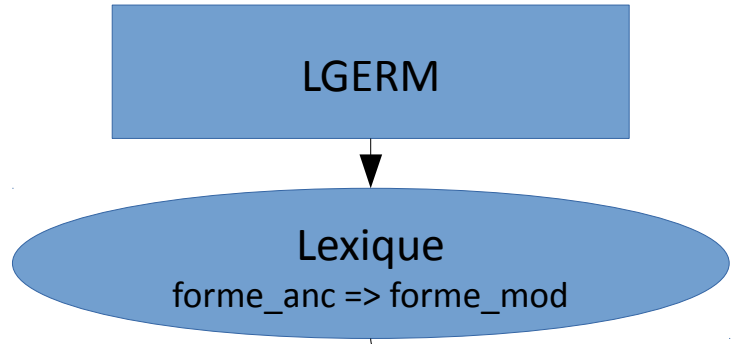


Désambiguïsation

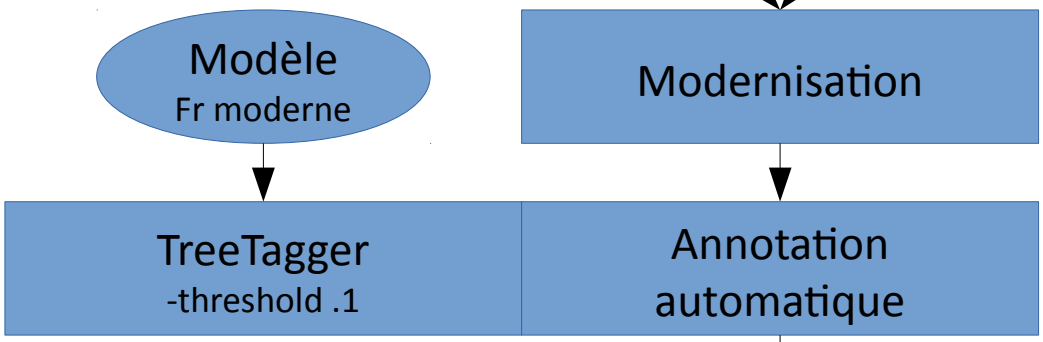



```
quelques        QUELQUE:Aq|QUELQUE:Di
remarques       REMARQUE:Nc|REMARQUER:Vvc
sur             SUR:Aq|SUR:Sp|SÛR:Aq|SÛR:R
les             LE:Da|LES:Pp|LÈS:Sp|LÉ:Nc
groupements     GROUPEMENT:Nc
```

Désambiguïsation

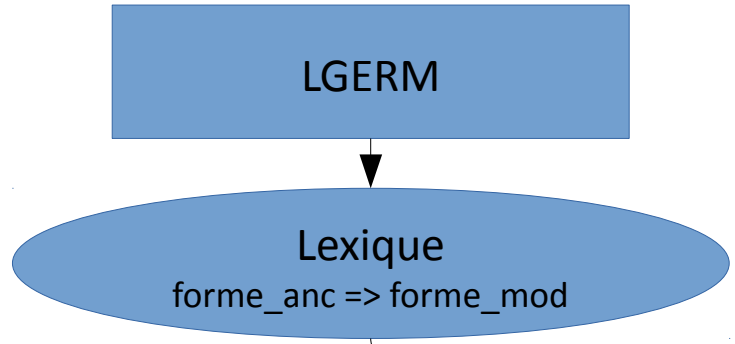


quelques	QUELQUE:Aq|QUELQUE:Di
remarques	REMARQUE:Nc|REMARQUER:Vvc
sur	SUR:Aq|SUR:Sp|SÛR:Aq|SÛR:R
les	LE:Da|LES:Pp|LÈS:Sp|LÉ:Nc
groupements	GROUPEMENT:Nc

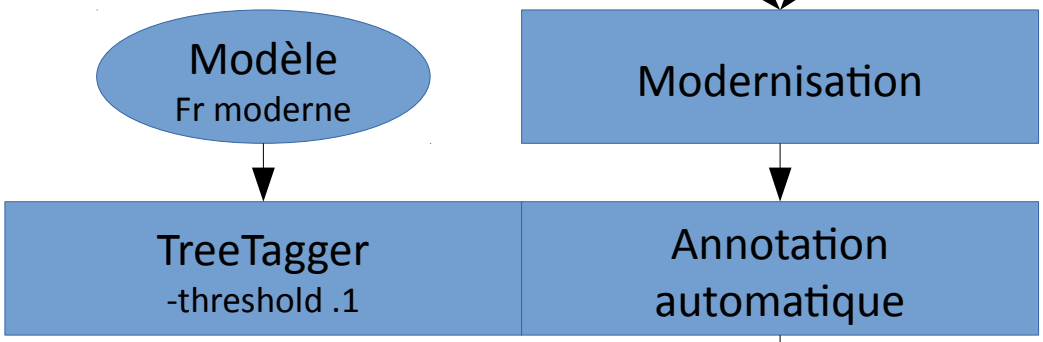


quelques	QUELQUE:ADJ|QUELQUE:DET
remarques	REMARQUE:NOM
sur	SUR:PRE
les	LE:DET
groupements	GROUPEMENT:NOM

Désambiguïsation

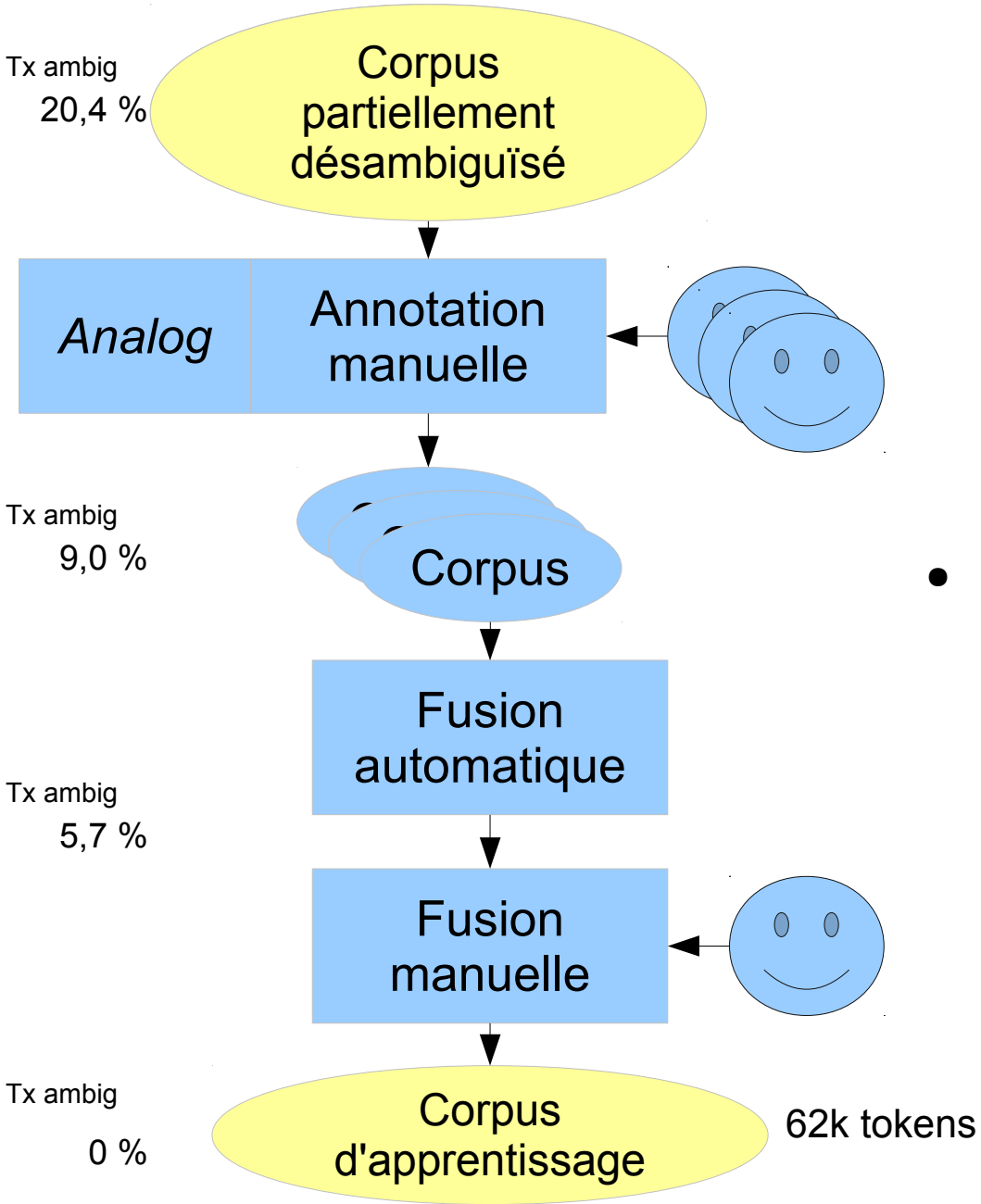



```
quelques        QUELQUE:Aq|QUELQUE:Di
remarques       REMARQUE:Nc|~~REMARQUER:Vvc~~
sur             ~~SUR:Aq~~|SUR:Sp|~~SÛR:Aq|SÛR:R~~
les             LE:Da|~~LES:Pp|LÈS:Sp|LÉ:Nc~~
groupements     GROUPEMENT:Nc
```



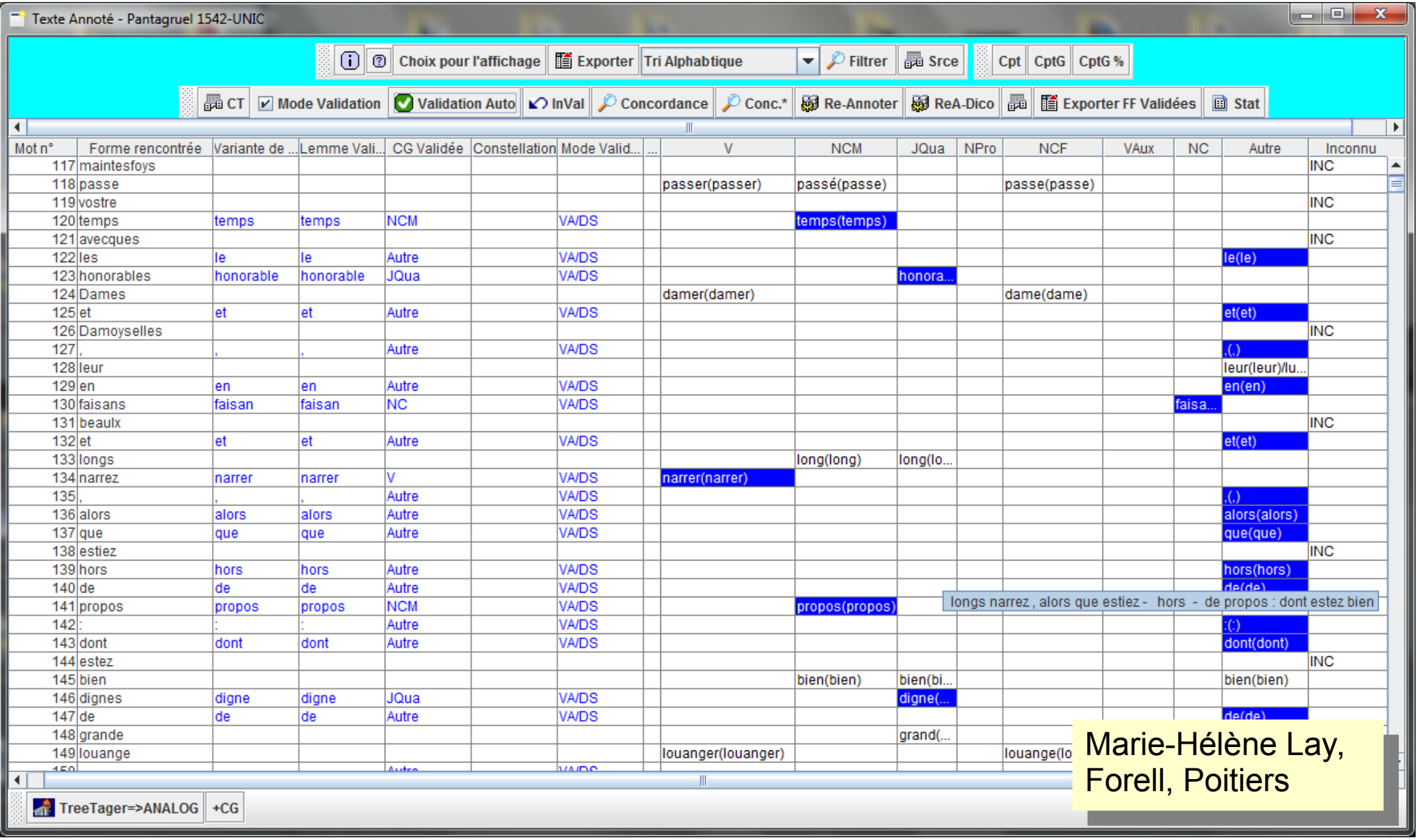

```
quelques        QUELQUE:ADJ|QUELQUE:DET
remarques       REMARQUE:NOM
sur             SUR:PRE
les             LE:DET
groupements     GROUPEMENT:NOM
```

Annotation manuelle et fusion



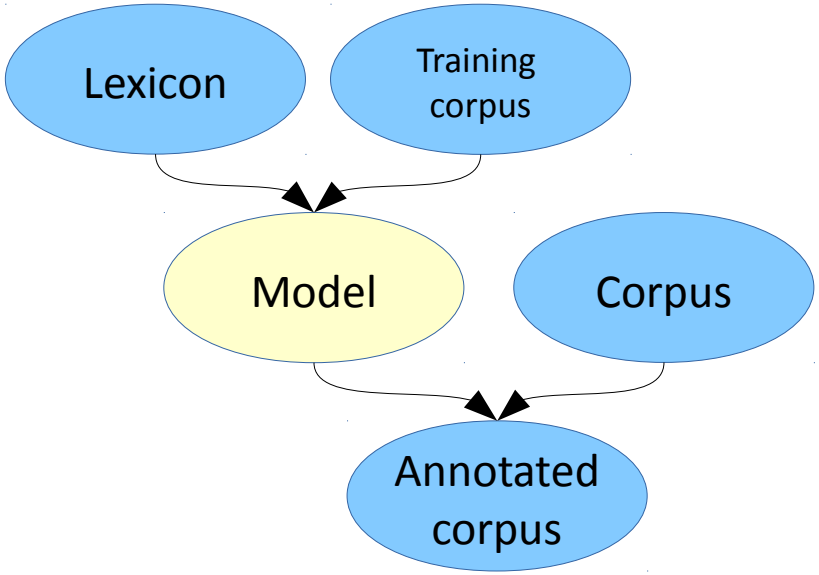
- Fusion automatique pour les cas « évidents » :
  - Au moins 2 annotateurs d'accord
  - Diacritiques

Manual annotation tool : *Analog*



Marie-Hélène Lay, Forell, Poitiers

A classic workflow

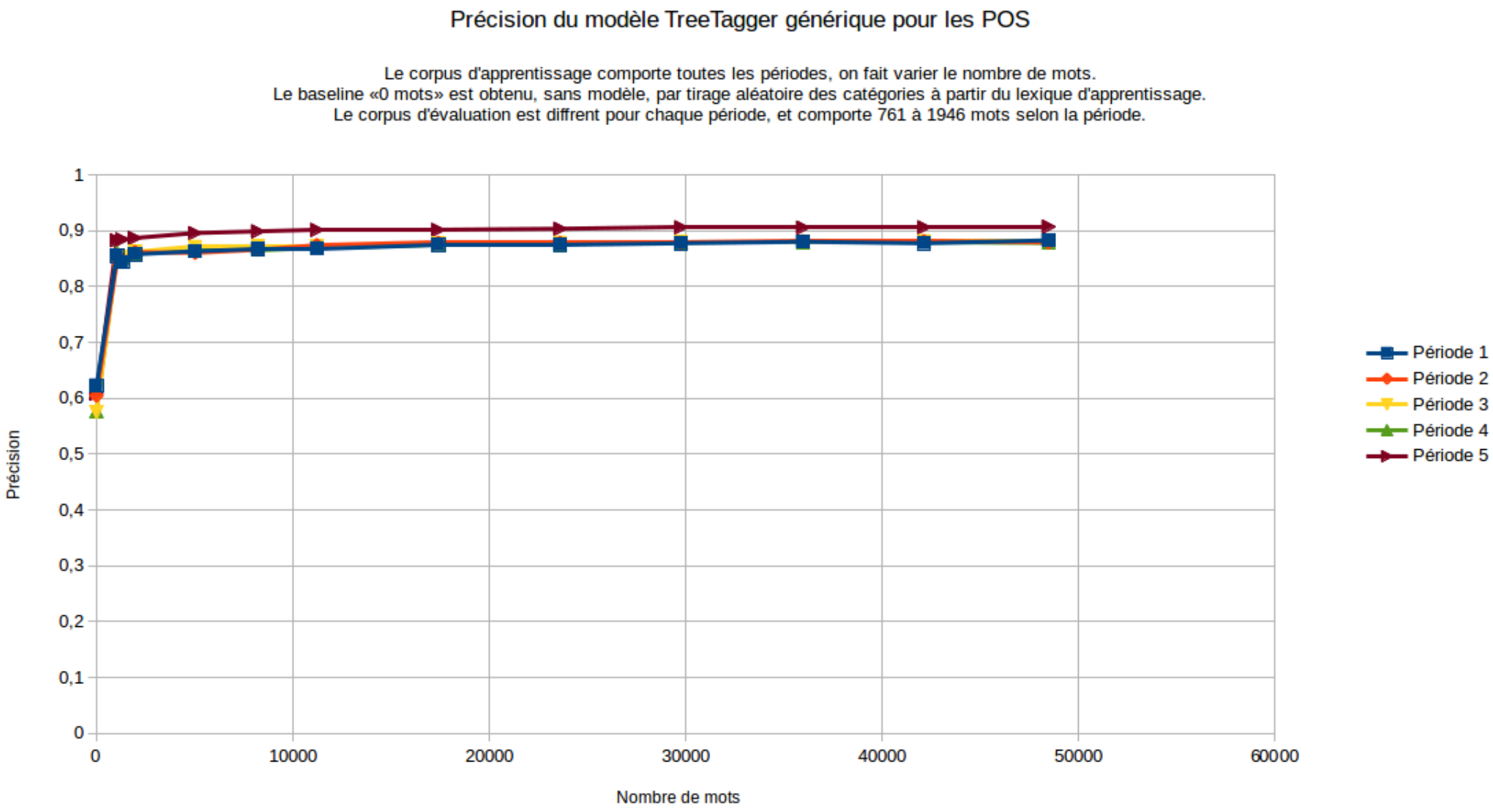


Tokens, POS, lemmas

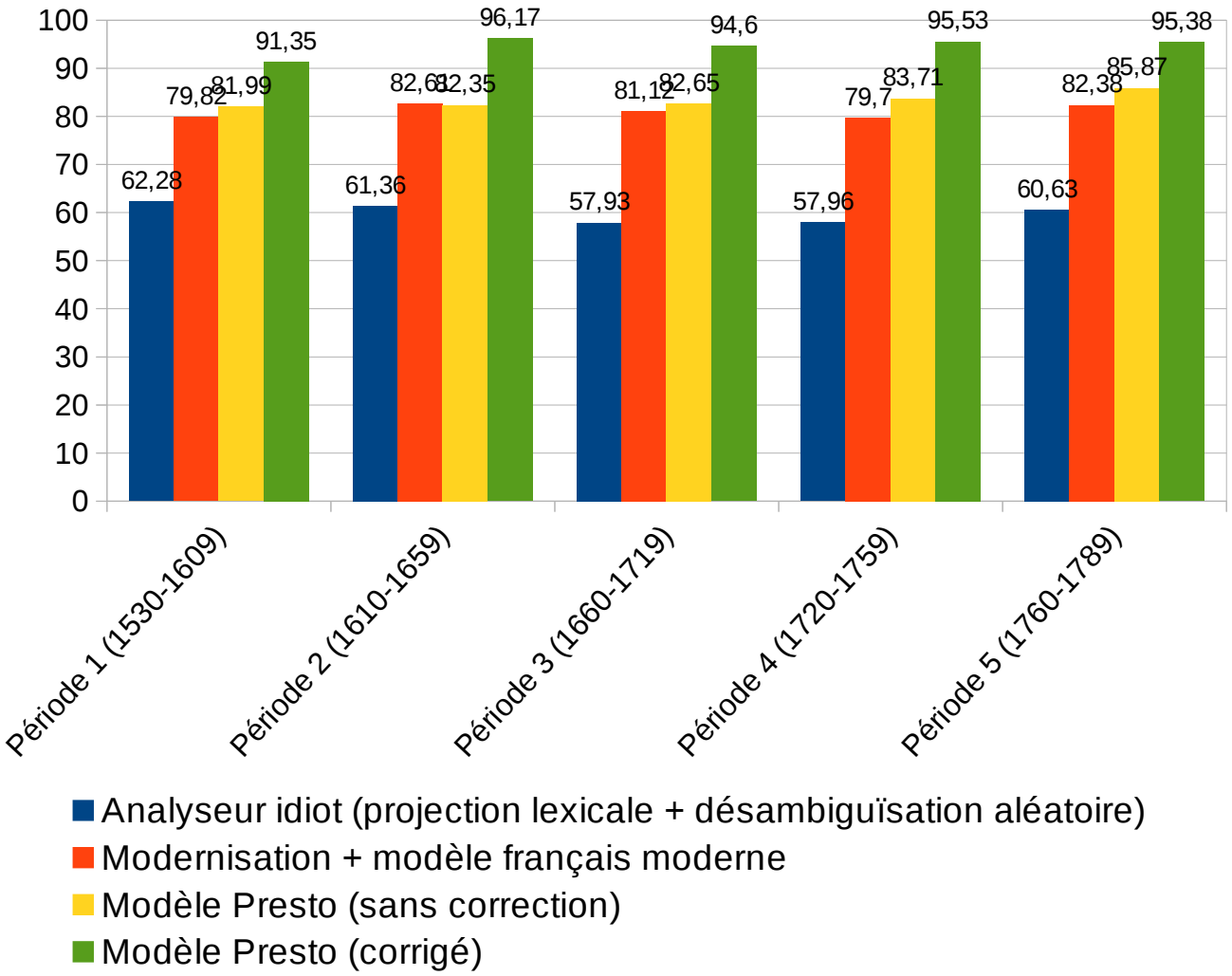
Model creation

- Training corpus divided in 3 parts
  - Train (80% − 49 630 tokens)
  - Dev (10% − 6 164 tokens)
  - Eval (10% − 6 110 tokens)
- *Autotuning* to find the best parameters for TreeTagger
  - cl 2 ; dtg 0,5 ; sw 1 ; ecw 0,06 ; atg 1,15
  - Precision gain : +0,05 %
- Evaluation
  - Train : 95,77 %
  - Dev : 94,28 %
  - **Eval : 94,46 %**

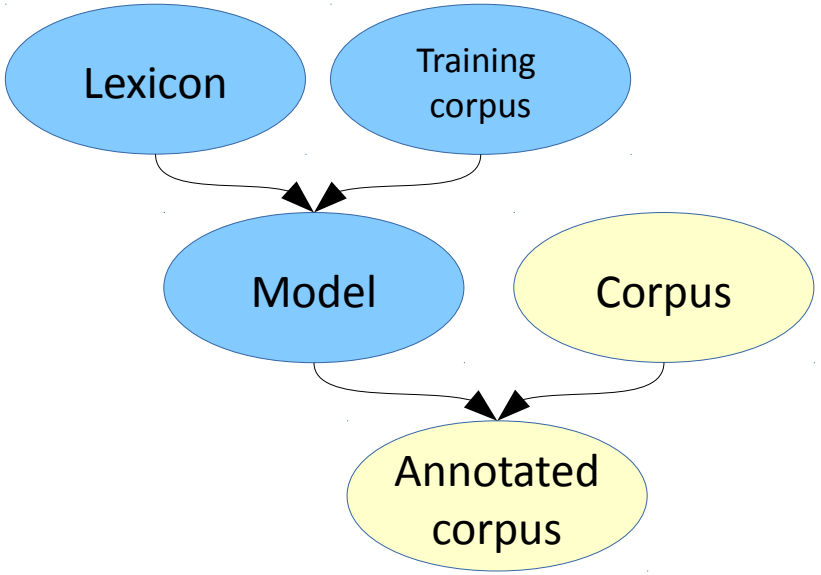
Model evaluation



Évaluation du modèle

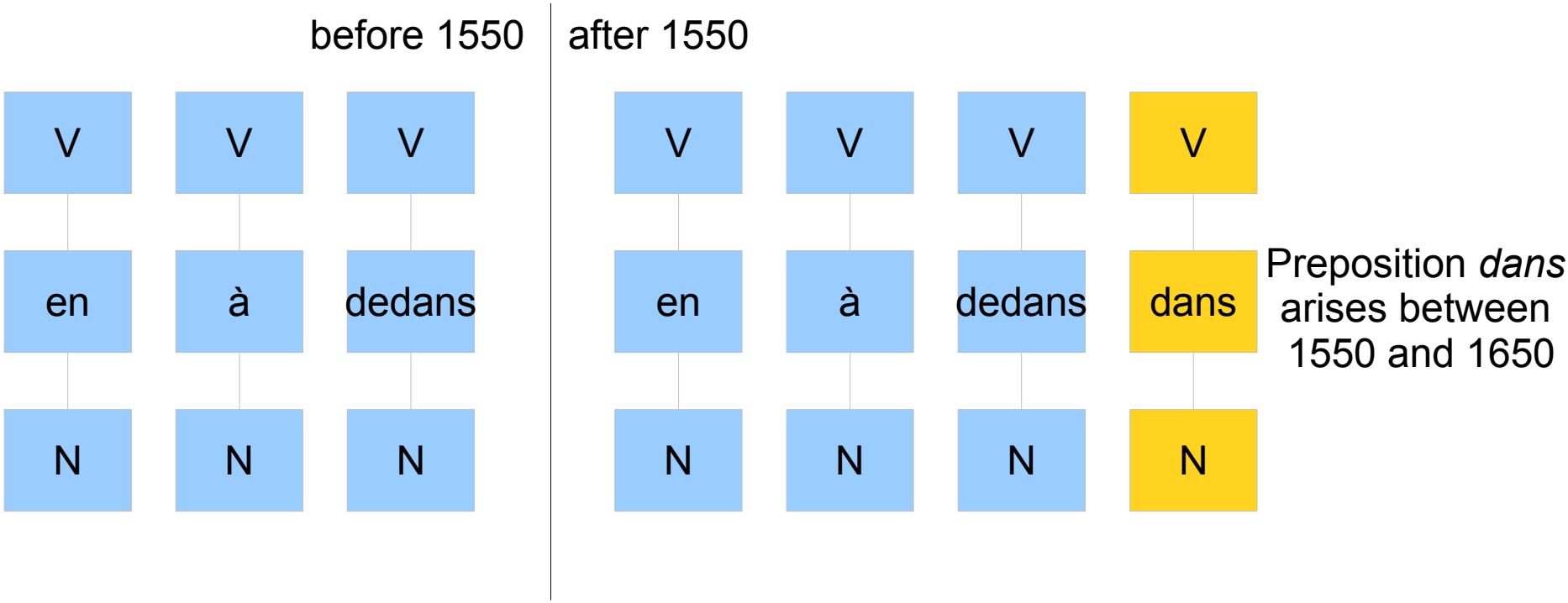


A classic workflow



Tokens, POS, lemmas

Prepositions dynamics



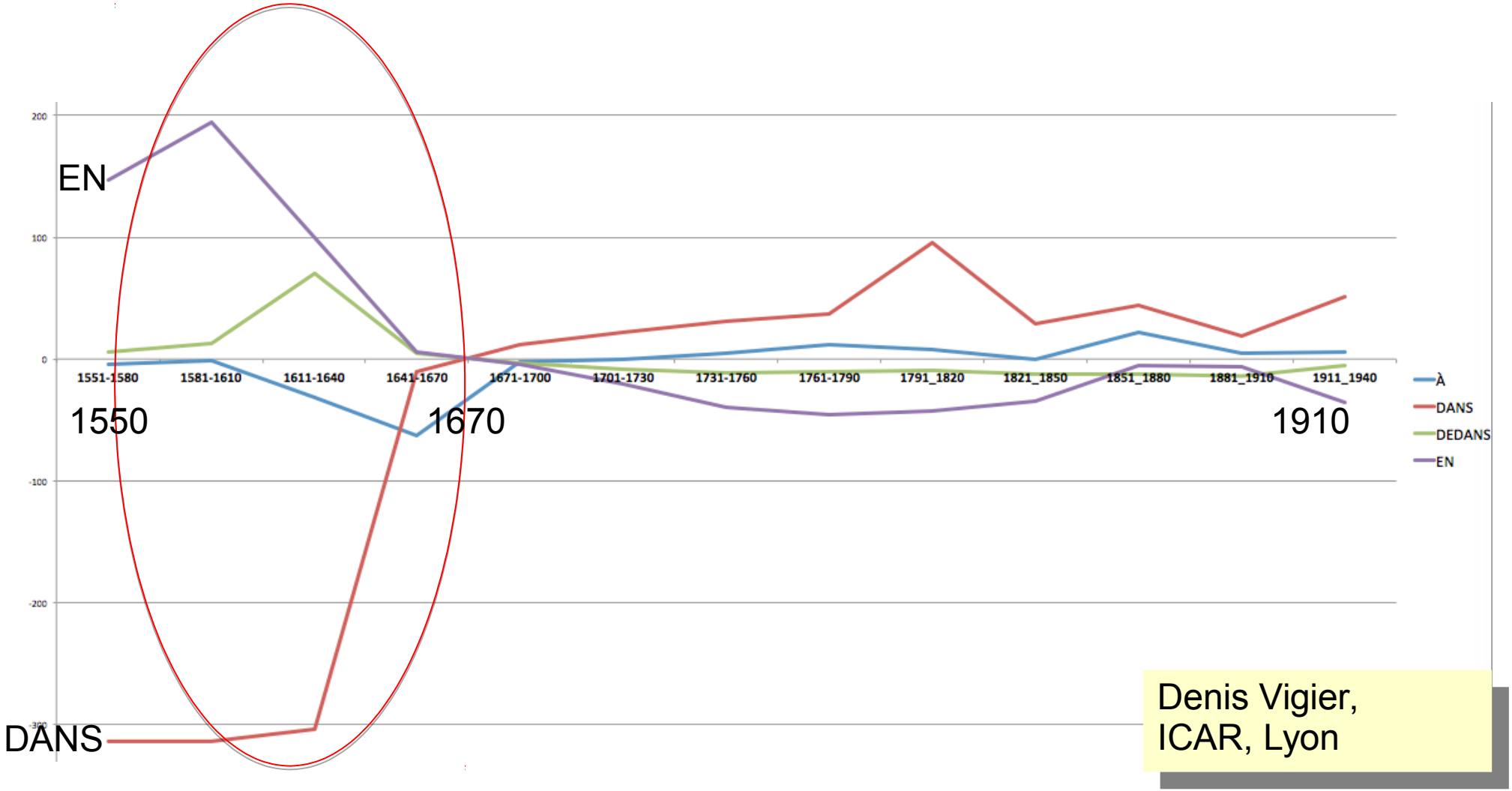


Diagramme 1. Évolution des scores de spécificité de Laffon affectés aux prépositions *en, dans*, *dedans* entre 1551 et 1900. Corpus Presto, partitionné en 13 tranches de 30 ans.

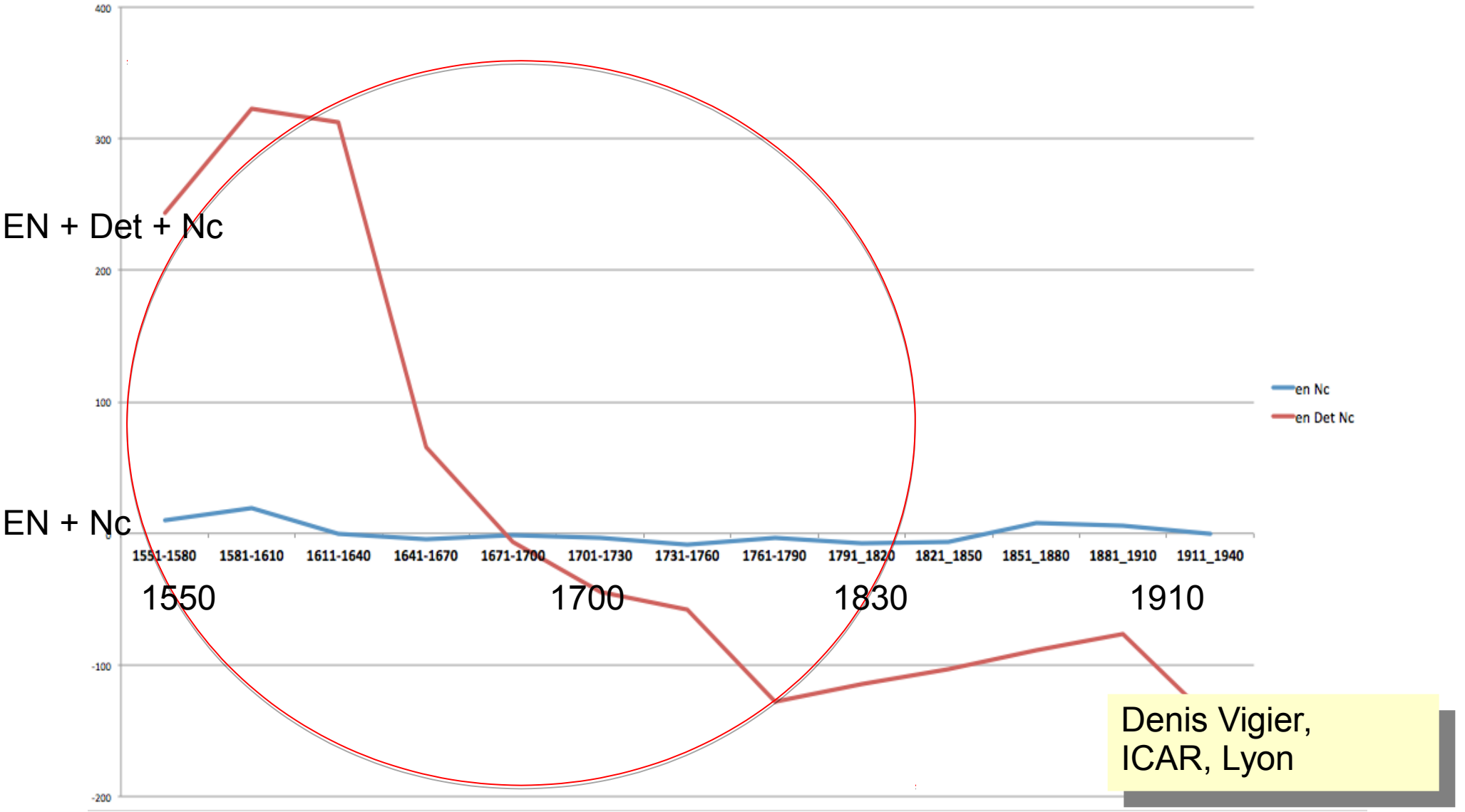


Diagramme 1. Évolution des scores de spécificité de Laffon affectés aux prépositions *en, dans*, *dedans* entre 1551 et 1900. Corpus Presto, partitionné en 13 tranches de 30 ans.

Encore plein de questions...

Problèmes

- Ambiguïté des lemmes
  - *Les fils d'Ariane*. → FILS, Nc *ou* FIL, Nc
  - *Congres* → CONGRÈS, Nc *ou* CONGRE, Nc
  - Il faudrait un TreeTagger de la WSD.
- Annotation panchronique
  - Comment couvrir différents états de la langue ?
  - Utiliser des jeux d'étiquettes différents ?
- Annotation en dépendances syntaxiques

1) Background

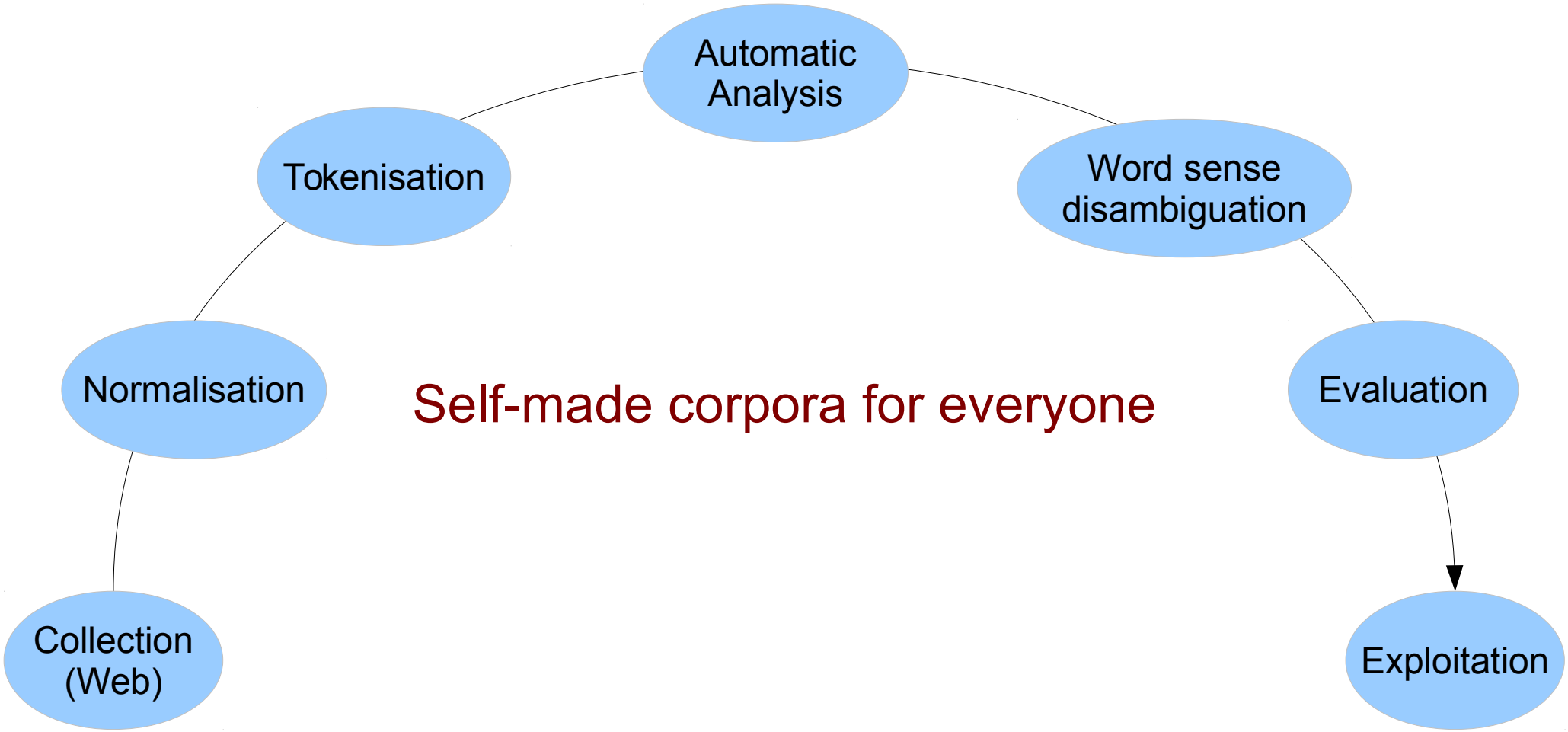
2) Past work

- Internet Relay Chat (*corpora by humans*)
- Scientific texts (*corpora for humans*)
- Corpora for teaching (*corpora for humans*)
- « Old » French texts (*corpora by humans*)

3) Future work

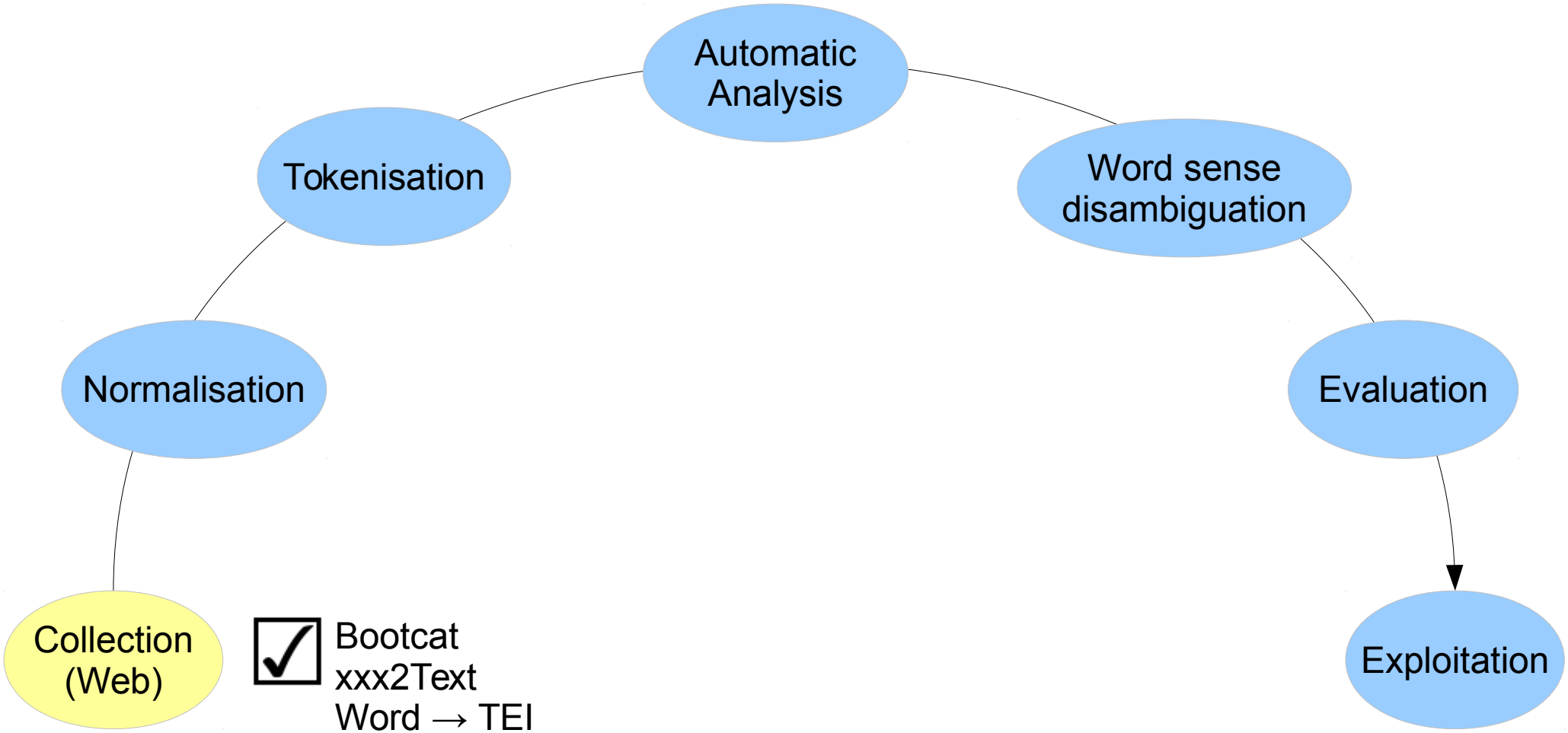
- Self-made corpora for everyone

Future work

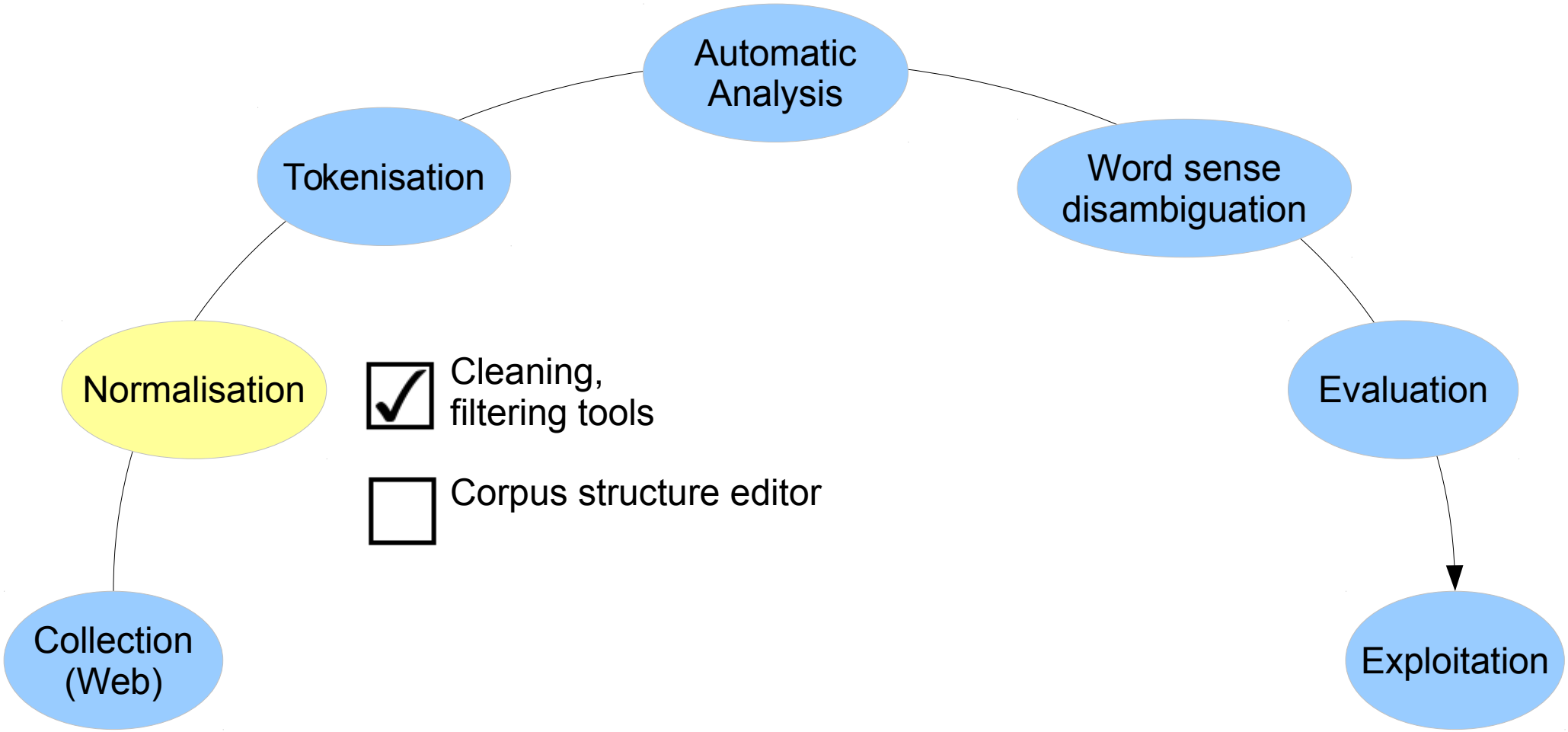


Self-made corpora for everyone

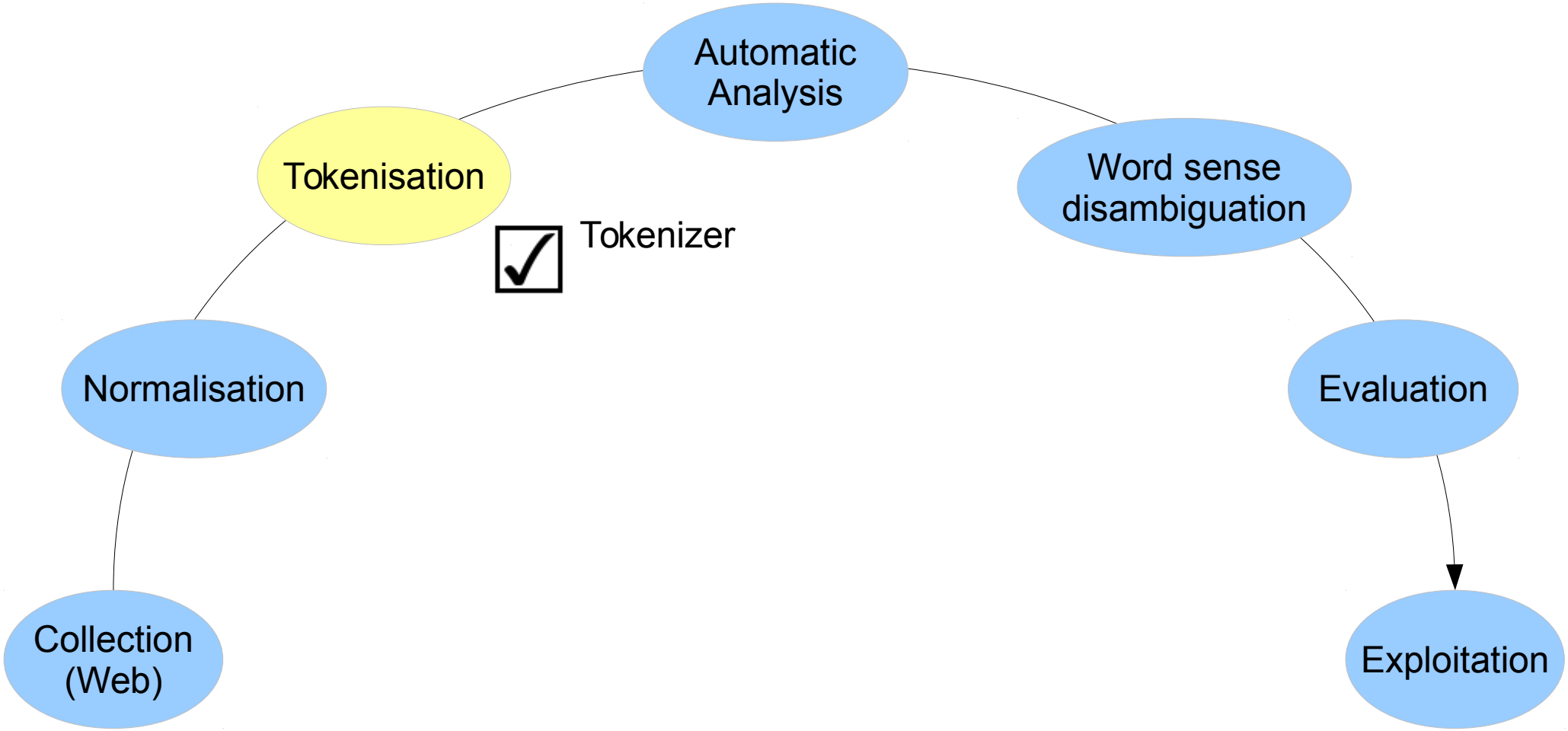
Future work


Automatic Analysis
Tokenisation
Word sense disambiguation
Normalisation
Evaluation
Collection (Web)
Bootcat
xxx2Text
Word → TEI
Exploitation

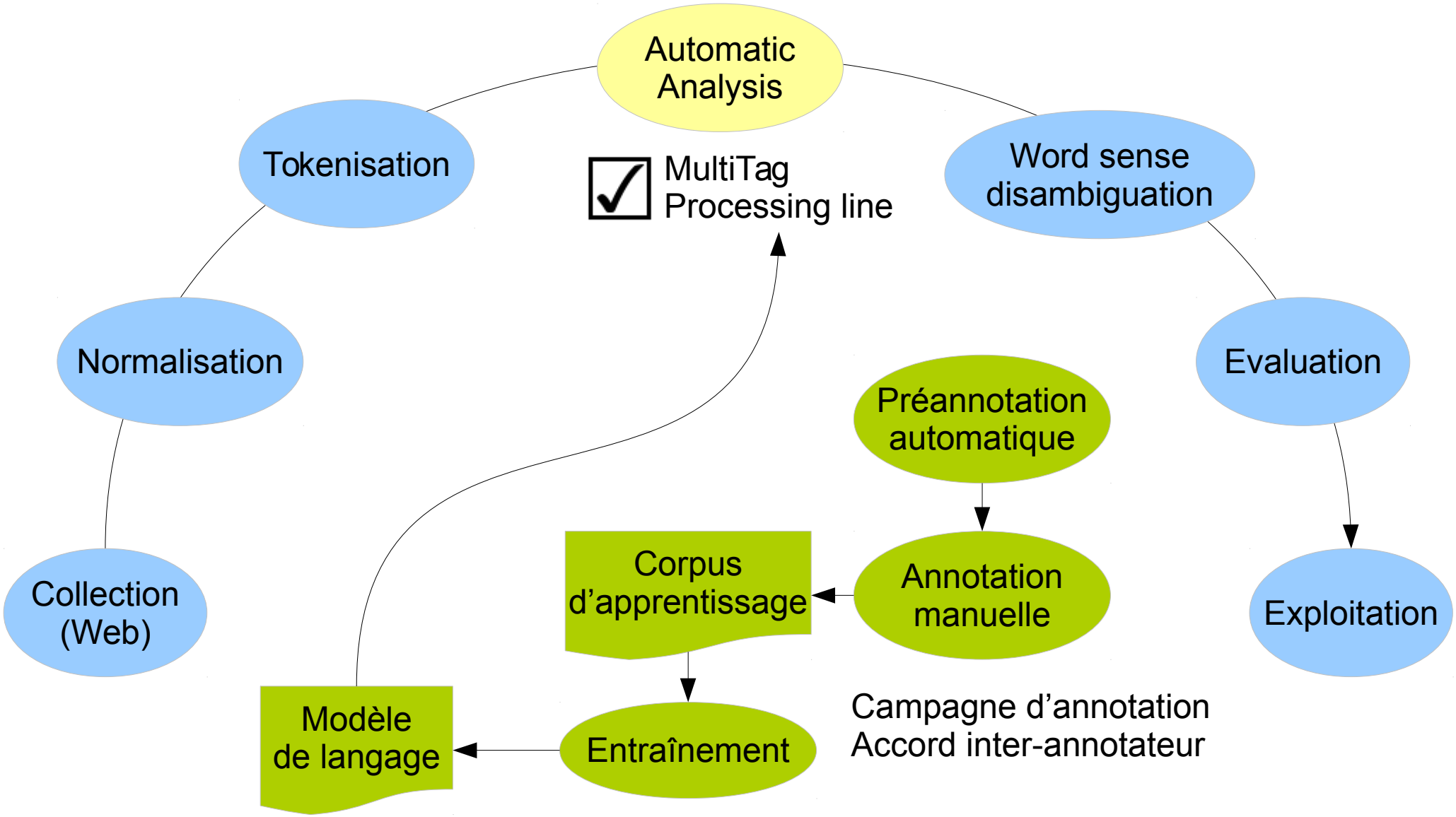
Future work


Automatic Analysis
Tokenisation
Word sense disambiguation
Normalisation
Evaluation
Cleaning, filtering tools
Corpus structure editor
Collection (Web)
Exploitation

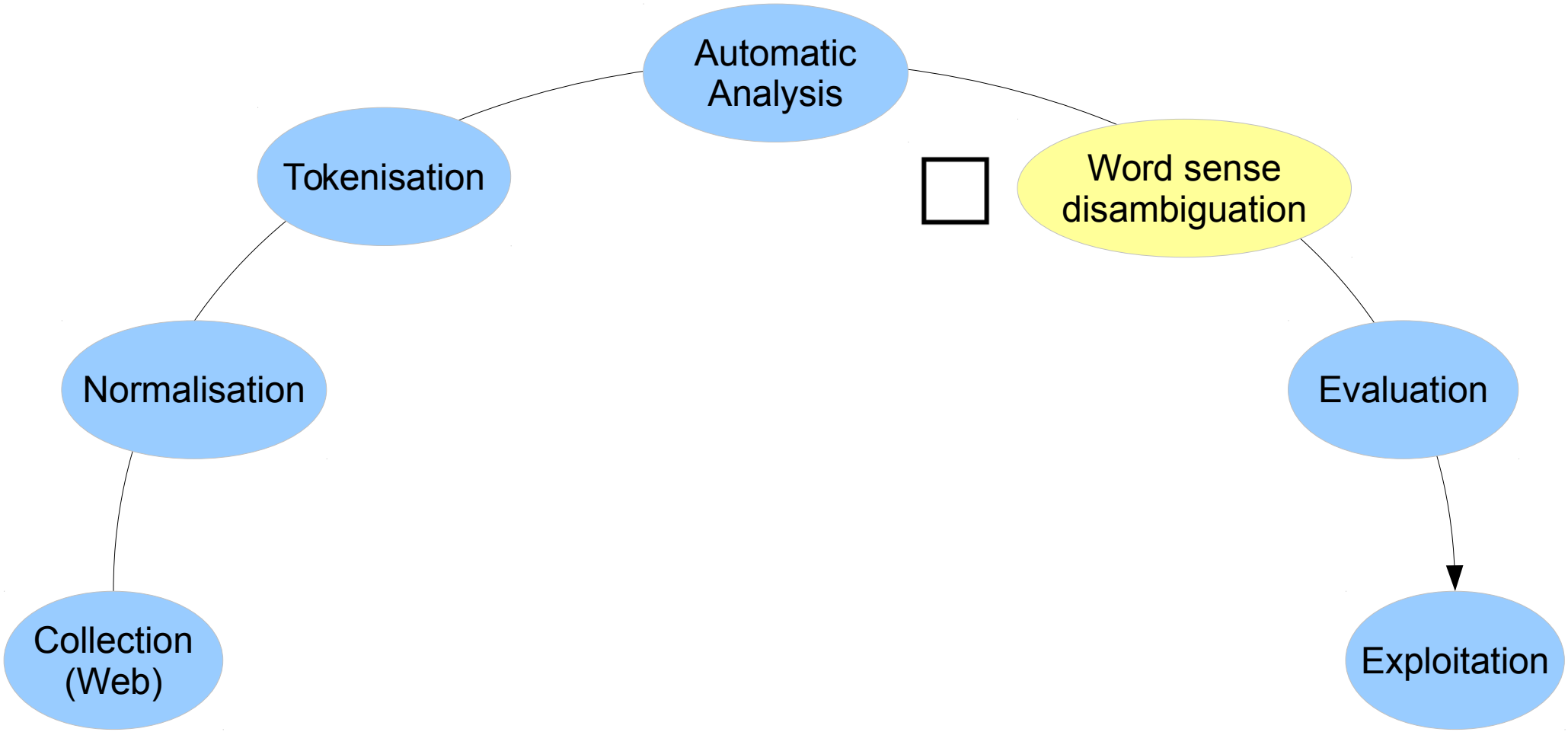
Future work


Automatic Analysis
Tokenisation
Tokenizer
Word sense disambiguation
Normalisation
Evaluation
Collection (Web)
Exploitation

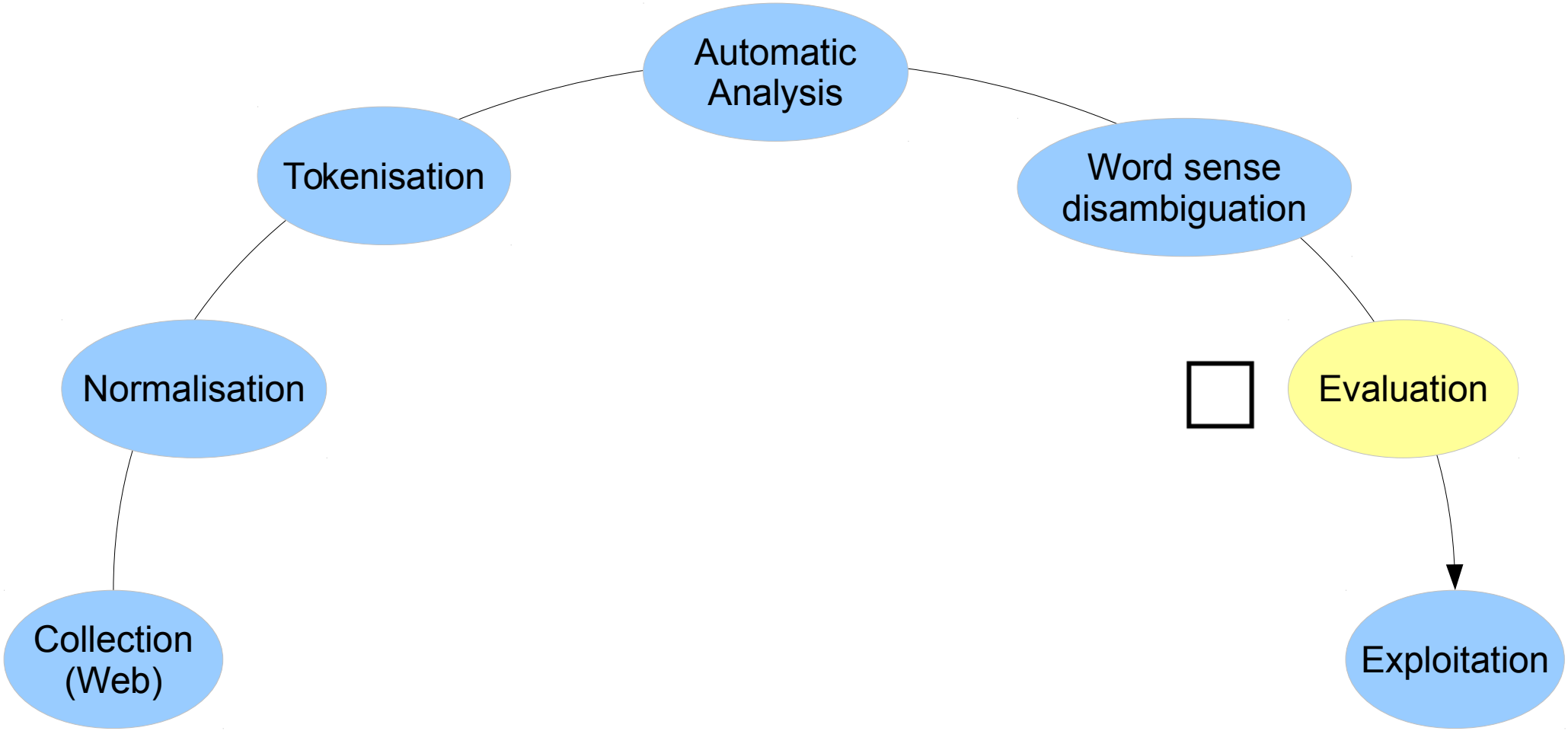
Future work


Automatic Analysis
Tokenisation
MultiTag Processing line
Word sense disambiguation
Normalisation
Evaluation
Préannotation automatique
Collection (Web)
Corpus d'apprentissage
Annotation manuelle
Exploitation
Modèle de langage
Entraînement
Campagne d'annotation
Accord inter-annotateur

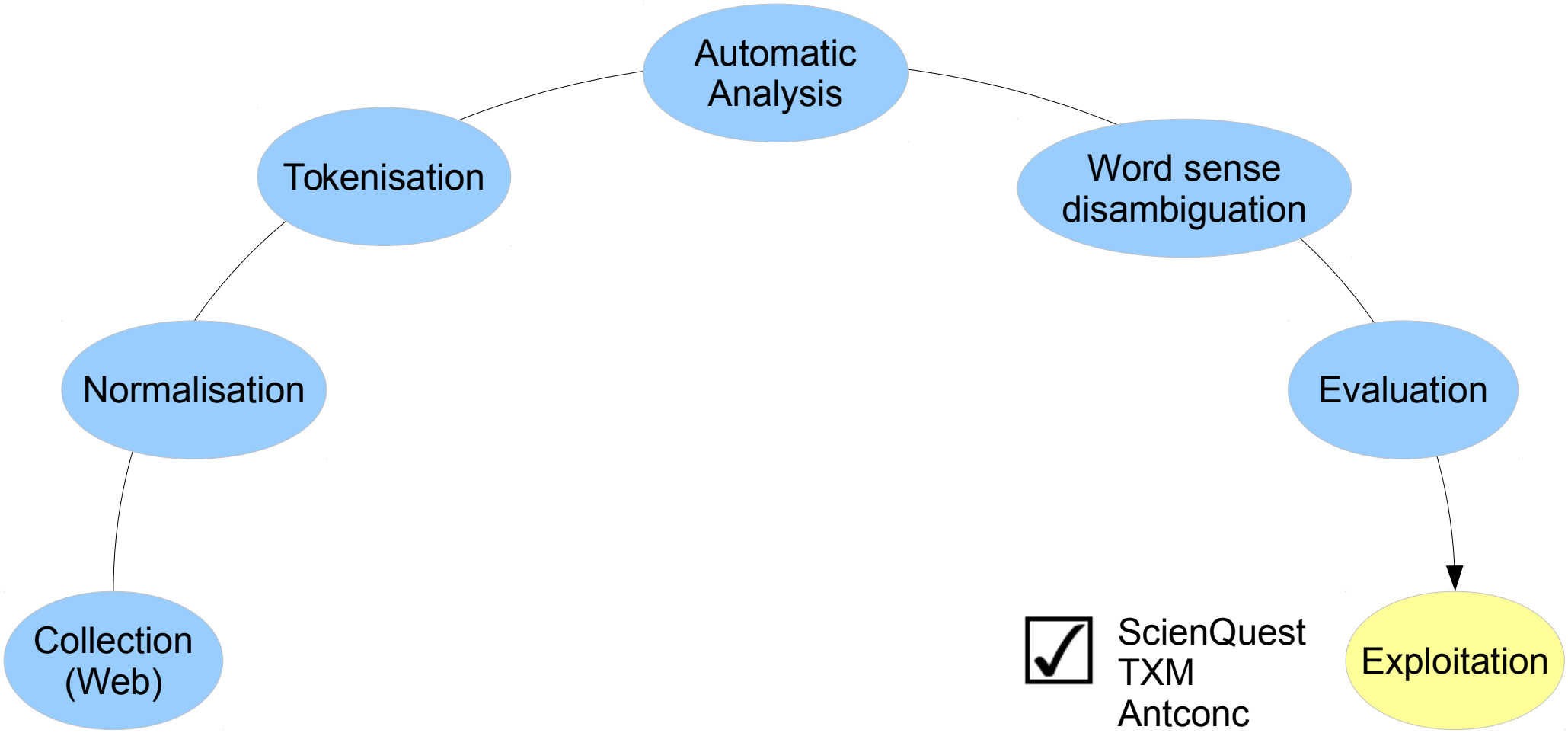
Future work


Automatic Analysis
Tokenisation
Word sense disambiguation
Normalisation
Evaluation
Collection (Web)
Exploitation

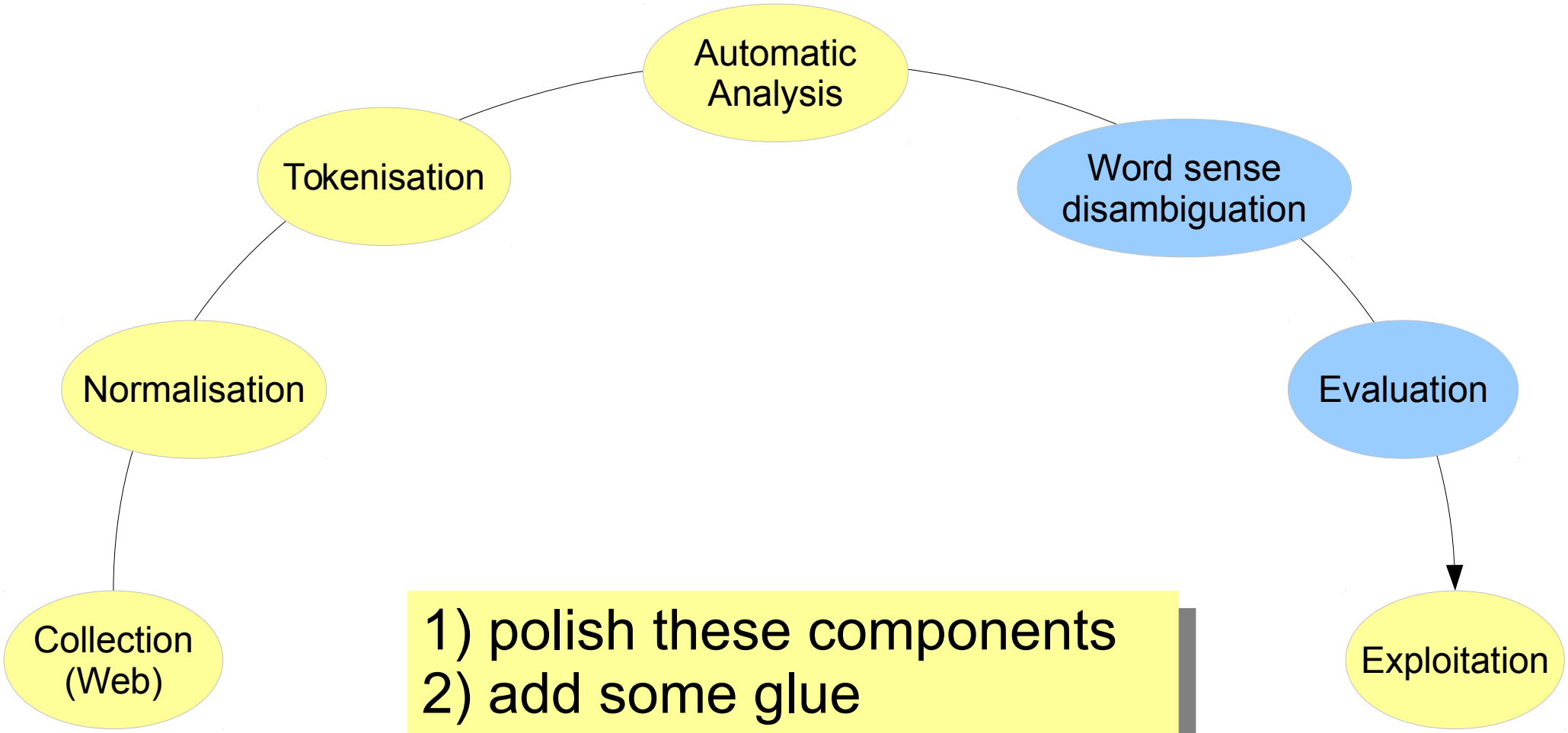
Future work


Automatic Analysis
Tokenisation
Word sense disambiguation
Normalisation
Evaluation
Collection (Web)
Exploitation

Future work


Automatic Analysis
Tokenisation
Word sense disambiguation
Normalisation
Evaluation
Collection (Web)
ScienQuest
TXM
Antconc
Exploitation

Future work



1) polish these components
2) add some glue
3) add a GUI
4) enjoy !

Conclusion

Ce qui m'intéresse

- Travailler avec des linguistes (mais pas que...)
- Passer à la moulinette de l'écrit (si possible bizarre)
- Créer des outils/ressources utilisables

Perspectives

- Accessibilité des corpus
- Outils qui vont voir au-delà de la phrase
  - dimension textuelle
  - dimension dialogique
- TAL en diachronie