

Université Côte d'Azur
UMR 7320 : Bases, Corpus, Langage
Séminaire
Logométrie. Corpus, Traitements, Modèles

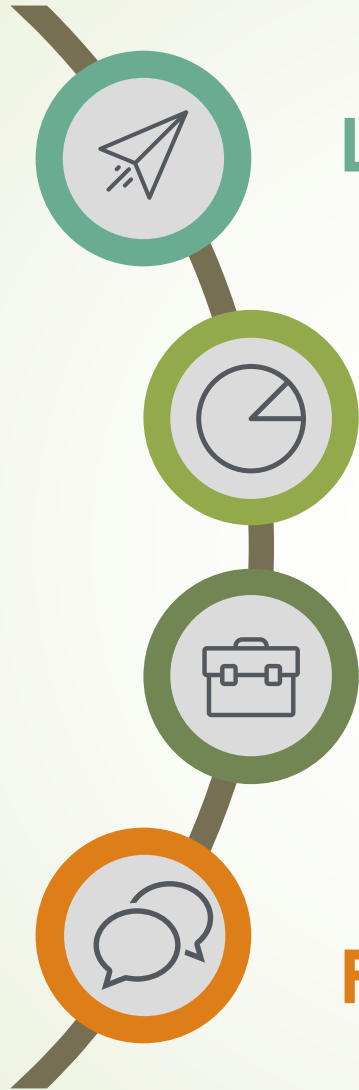
Corpus numériques en FLE

Cristelle CAVALLA

6 mai 2021

**Sorbonne
Nouvelle**  DILTEC - EA 2288
Didactique des langues,
des textes et des cultures

PRÉSENTATION

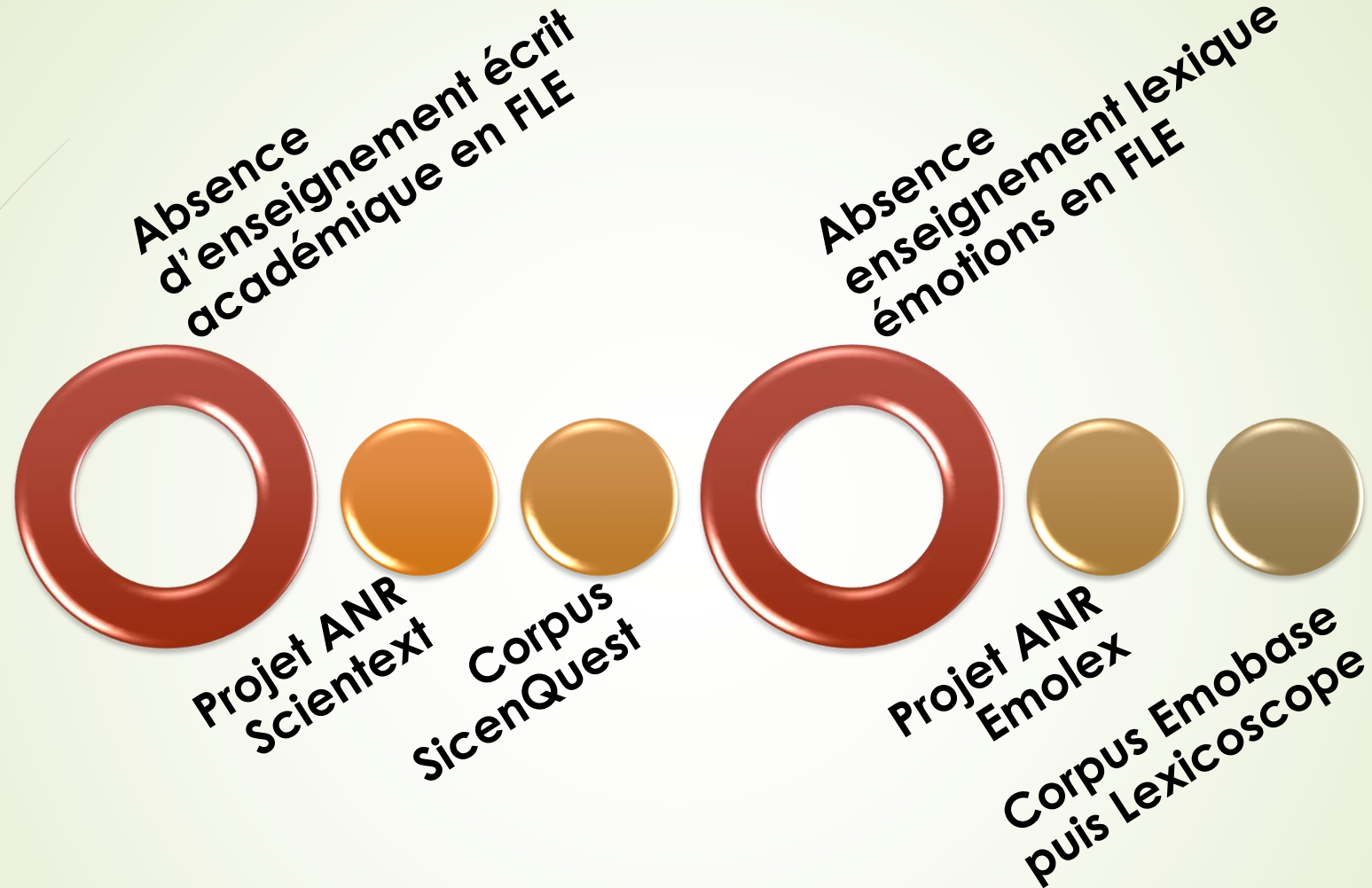


LES CORPUS EN CLASSE DE LANGUE

CORPUS ET ÉCRITS SCIENTIFIQUES

CORPUS ET ÉMOTIONS

PERSPECTIVES

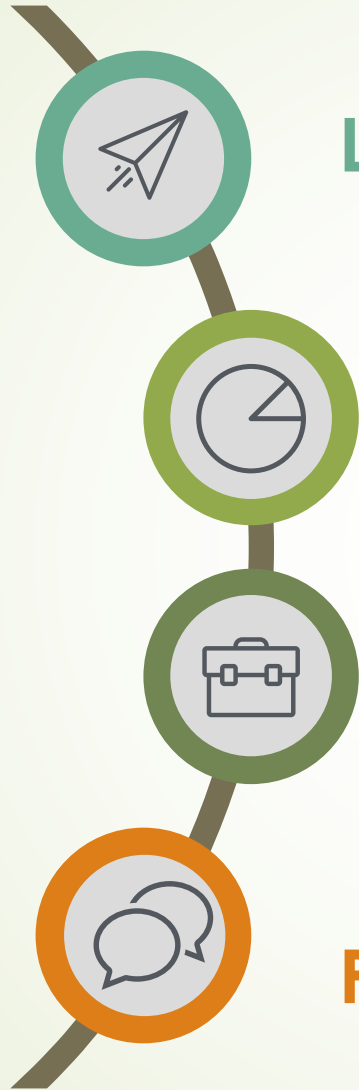


Cavalla C. Lexique transdisciplinaire et enseignement aux étudiants allophones. A. Tutin; M-P. Jacques. *Lexique transversal et formules discursives des sciences humaines*, ISTE Editions, pp.191-214, 2018.

Cavalla C. Exemple d'enseignement de la phraséologie transdisciplinaire à l'aide de corpus numériques en FLE. *La lettre de l'AIRDF*, Association internationale de recherche en didactique du français, 2018, La didactique du lexique, pp.43-47. <hal-02087791>

Cavalla C. Quel lexique pour quelles émotions en classe de FLE ?. *Le Langage et l'Homme*, EME éditions / L'Harmattan, 2015. Affects et acquisition des langues, L.2 (50/2), pp.115-128. <hal-01375964>

PRÉSENTATION



LES CORPUS EN CLASSE DE LANGUE

CORPUS ET ÉCRITS SCIENTIFIQUES

CORPUS ET ÉMOTIONS

PERSPECTIVES

Objectifs pour l'apprenant

Être capable
de rédiger un
écrit
académique

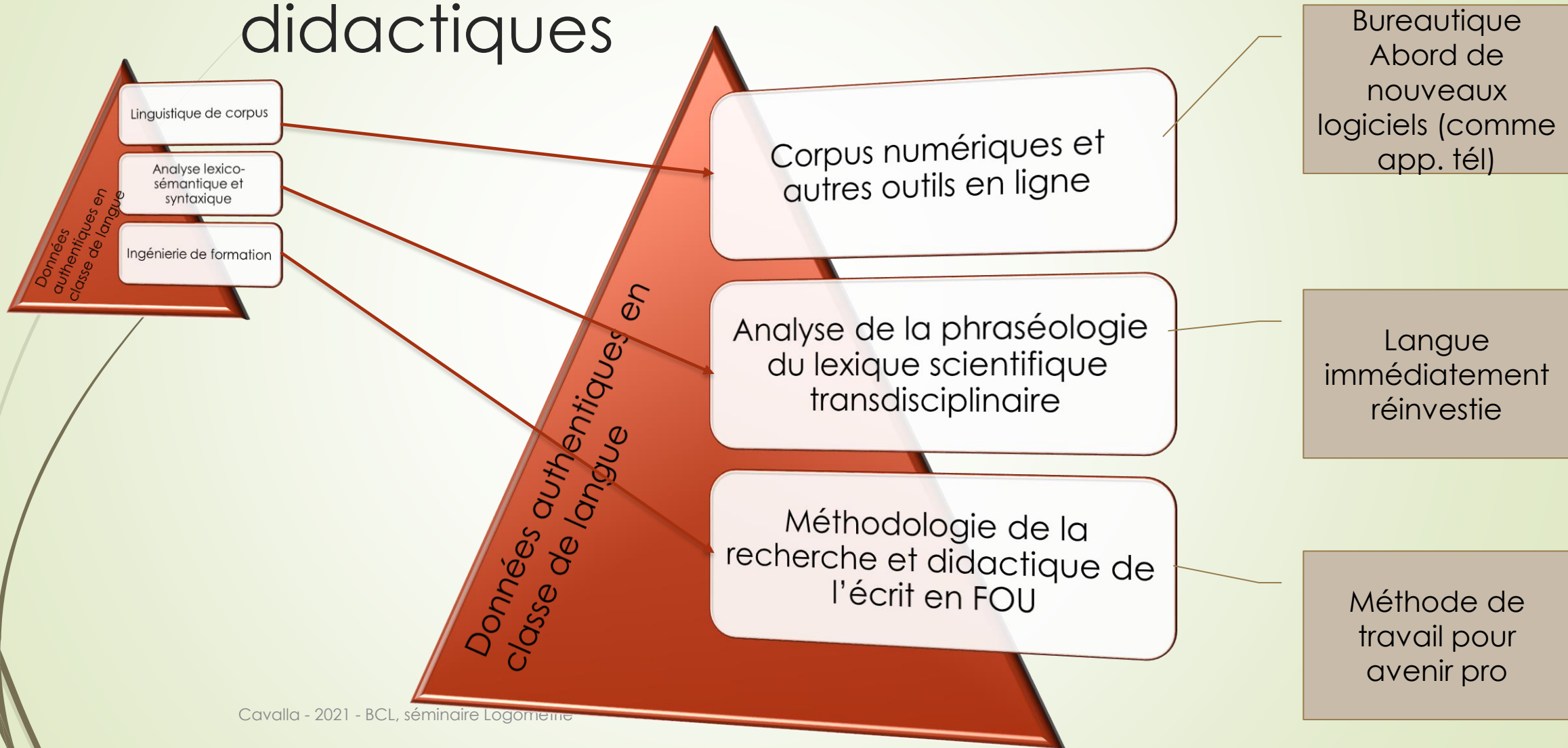


Être capable
d'utiliser des
corpus
numériques
pour rédiger

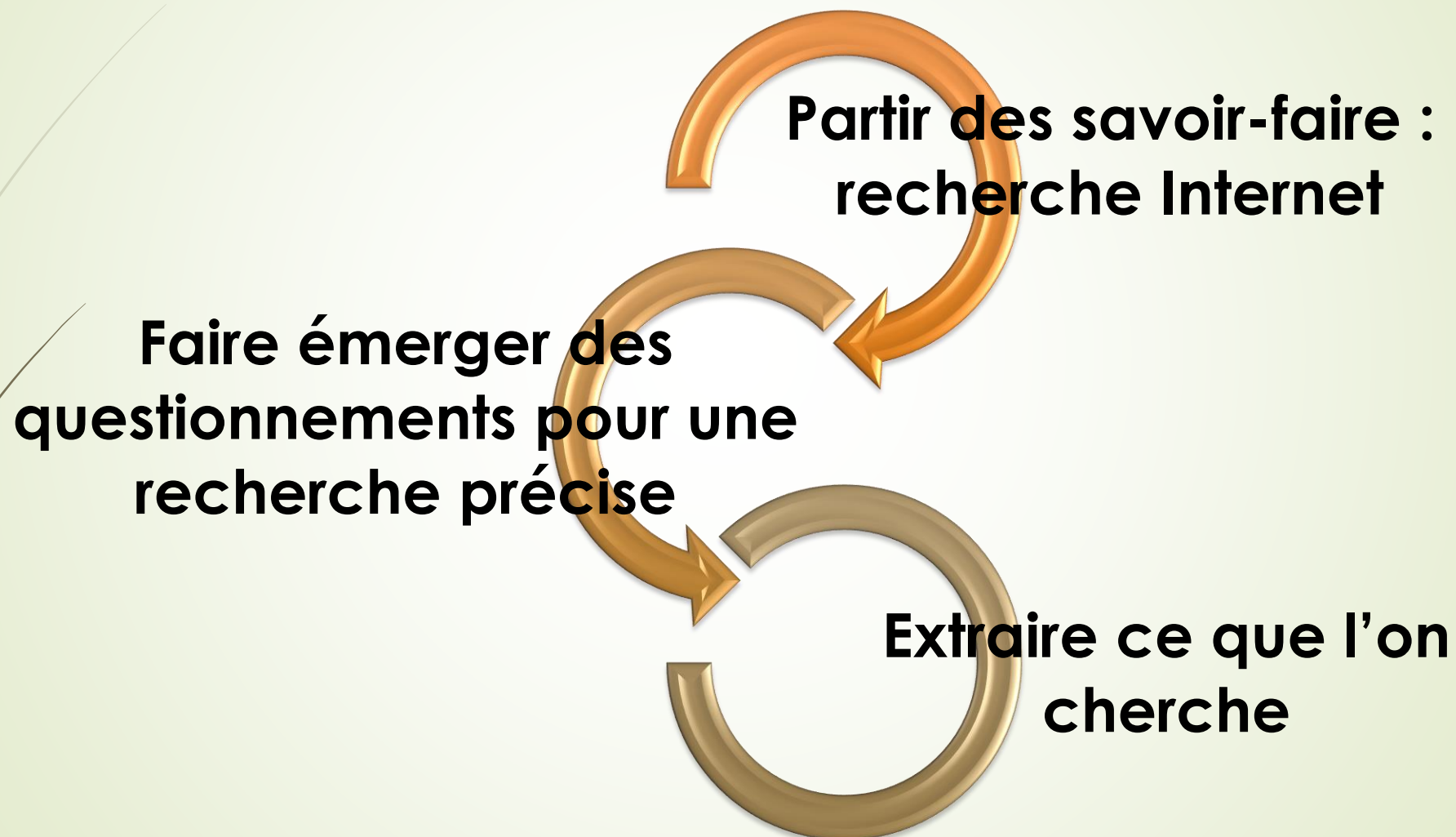


Être capable
d'utiliser les
éléments
linguistiques
appropriés à
cet écrit

Outils pour l'élaboration de séquences didactiques



Comment faire découvrir ?



Approche pédagogique

Objectif : autonomie

■ Directe : leurs questions

- *comment énoncer ses questions de recherche / ses hypothèses ? Comment présenter son sujet ?...*

■ Approche inductive : leurs besoins

- *Prise de conscience du contenu de l'écrit académique*
- *Organisation du contenu*
- *Recherche des éléments pour énoncer le contenu*

■ Onomasiologie : vers autonomie

- *Entrées sémantiques*

Approche directe

Objectif : spécifier un élément linguistique

■ Indirecte : arrêt sur image

- *concordancier sur un seul élément ou sur une classe sémantique*

■ Comprendre l'utilisation : autonomie de la méthode d'interrogation

épistémologique (comment fonctionne cette langue ?)

Approche indirecte

Approche Directe : recherche sur outils connus



Dans quel type de phrases et de textes trouve-t-on l'expression « en particulier » ?



Comment sélectionner les exemples trouvés ?
Comment vérifier les contextes ?

...

Approche Directe

1^{er} corpus : des chiffres

CORPORA COLLECTION

Corpus

French mixed corpus based on material from 2012
Sentences: 74,823,426 · Types: 7,873,935 · Tokens: 1,468,766,604 more...

Word: **Corpus** Number of occurrences: 1,221 Rank: 50,231 Frequency class: 16

See also: corpus, CORPUS

Similarity based on Cooccurrences: Corpus Christi | Christi

Examples: S'inscrivant dans le cadre du festival Corpus,...

Cooccurrences:

Corpus Christi (6,165), Christi (5,196), Domini (1,353), Texas (864), Habeas (798), Hermeticum (556), Corpus hippocratique (486), l'Habeas (486), corpus (418), hippocratique (369), Verum (323), ((315), textes (313), Corpus Christi College (309), Juris (299), Juris (299), Christianorum (287), Corpus Delicti (287), Corpus Christianorum (287), Delicti (287),) (279), inscriptionum (265), Makina (264), du (251), inscriptionum (236), Corpus juris civilis (221), civilis (205), Corpus inscriptionum latinarum (199), Ave Verum Corpus (199), Galveston (197), latinarum (192), solennité (192), procession (182), Jérôme Prieur (179), Fr (153), Justinien (152), Ave (141), Cambridge (133), d'Habeas (121), Langage (120), Rhésus (117), : (117), Andrea Bocelli (116), Prieur (112), Bocelli (111), S (109), College (102), Medicorum (98), Etampois (97), langue (96), Linguistiques (95), Latinarum (88), Separatum (88)

Christi : 5 196 occurrences

CORPORA COLLECTION

corpus

French mixed corpus based on material from 2012
Sentences: 74,823,426 · Types: 7,873,935 · Tokens: 1,468,766,604 more...

Word: **corpus** Number of occurrences: 6,108 Rank: 17,112 Frequency class: 14

See also: Corpus, CORPUS

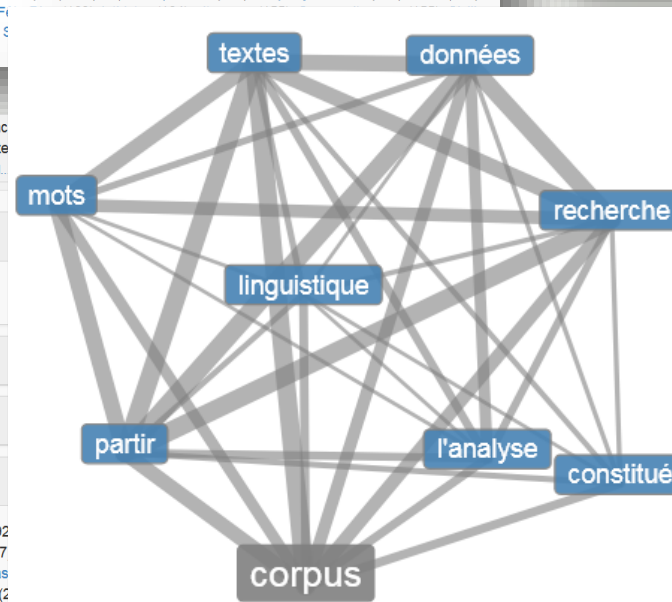
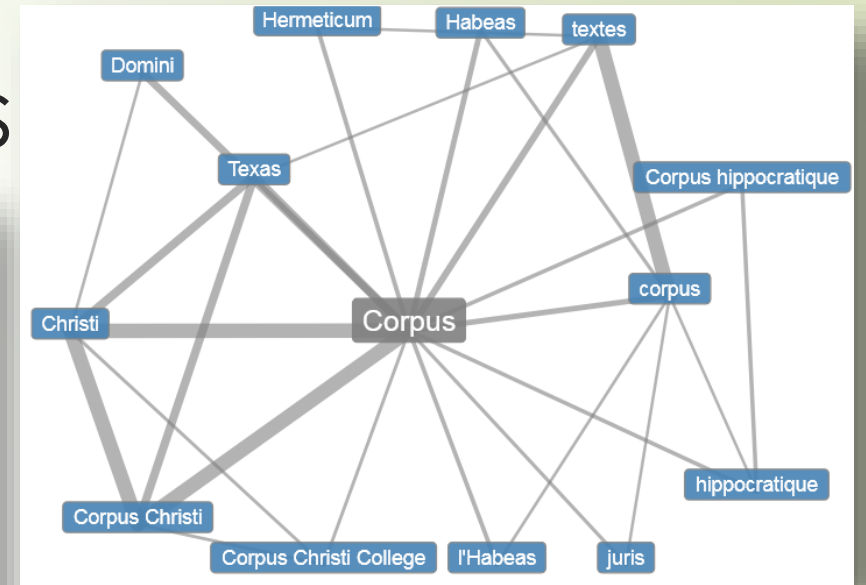
Similarity based on Cooccurrences: lexique | recueil | vocabulaire | catalogue | répertoire

Examples: Le corpus documentaire réuni pour cette étude...

Cooccurrences:

textes (2,863), l'habeas (1,539), habeas (1,294), d'habeas (941), partir (707), de (702), linguistique (675), mots (651), un (648), d'un (648), des (592), idéologique (448), connaissances (434), théorique (426), Corpus (418), linguistiques (411), oraux (398), et (390), ((357), recherches (350), du (347), (309), constituer (306), documents (300), littéraires (289), outils (280), textuels (266), dans (266), sur (265), langues (260), d'analyse (259), Habeas littérature (233), T-LAB (231), législatif (230), automatique (229), documentaire (226), l'ensemble (225), différents (220), L'habeas (220), lexicales (2

textes : 2863 occurrences



Découverte des extractions :
Comment les lire ?

Approche Directe : 2^e corpus : un concordancier

Lemma : *considérer*

Observer à droite : *comme / que*

Compleat Lexical Tutor

Home > Concordancers > French Input [«Back](#) (Back keeps original settings) Copiable extract-Link to this data >> [here](#)

Concordance for lemma *considérer* in Fr_le_monde.txt sorted 1 wd left of key Dictionnaire Fren_Eng Speak Fr-F

Extract: All | Q | Any 10 | 20 | 30 | 50

lemma *considérer* | Le Monde (1998) 1 110 392 sorted 1 wd left +assoc on left

10. tre forme de procès eût été le pire des **affronts**. [CONSIDÉRANT](#) que la meilleure défense était l'attaq

11. -il poursuivi "Le Pakistan, a-t-il enfin **ajouté**, [CONSIDÈRE](#) que ces deux points doivent être discuté

12. CDU et le SPD, un scénario que 40 % des **Allemands** [CONSIDÈRENT](#) comme le plus probable à huit mois des

13. e tribunal administratif. Les juges avaient **alors** [CONSIDÉRÉ](#) que l'on ne pouvait fixer "de conditions

14. c à Alcatel à amplifier son rebond. Les **analystes** [CONSIDÈRENT](#), de façon générale, que le groupe en a

15. ssi lourdes, reconnaît Compaq. Certains **analystes** [CONSIDÈRENT](#) que la remise en confiance de la clien

16. ux seules ventes de logiciels. Certains **analystes** [CONSIDÈRENT](#) qu'en développant une activité de cons

17. au Collège de France, âgé de soixante-quinze **ans**, [CONSIDÉRÉ](#) comme l'un des plus grands économistes f

18. "affaires générales" d'Elf-Aquitaine, **aujourd'hui** [CONSIDÉRÉ](#) comme le personnage central de l'affaire

19. , Charlotte Perriand créa des meubles **aujourd'hui** [CONSIDÉRÉS](#) comme des classiques de la modernité. E

20. dent le relèvement ou la grâce, et quinze **autres**, [CONSIDÉRÉES](#) comme ne remplissant pas les critères

21. bunal impartial", les juges de Strasbourg **avaient** [CONSIDÉRÉ](#) que la cour d'assises n'avait pas procé

22. uption, la haute juridiction administrative **avait** [CONSIDÉRÉ](#) que le contrat conclu par le maire avec

23. de l'ancien ministre GATTEGNO HERVE **Un** [avocate](#), [CONSIDÉRÉE](#) comme une amie proche de Roland Dumas,

24. ri Philippi, Louis Schweitzer et Patrick **Baudry**), [CONSIDÈRE](#) que l'information judiciaire ne peut être

25. es moins "rébarbatives", comme en Staps. M. **Borel** [CONSIDÈRE](#) enfin que les sciences sont actuellement

26. rovenance de Londres. Le gouvernement **britannique** [CONSIDÈRE](#) comme superfétatoire l'adoption d'une lo

27. France n'est pas davantage pour la **centralisation** [CONSIDÉRÉE](#) comme une sorte de dogme. Quant à l'All

28. oire le fait qu'il y a moins de cent ans la **Chine** [CONSIDÉRAIT](#) comme relevant de sa souveraineté la p

29. e l'écrivain silencieux". La thèse de Marc **Comina** [CONSIDÈRE](#) les textes de l'écrivain et ceux des cri

30. se trouve dans une situation juridique **complexe**, [CONSIDÉRÉE](#) par Me Schoenbach comme une "injustice

31. à éviter, comme les grands conglomérats **coréens**. [CONSIDÉRÉS](#) à tort comme une source de réussite, il

32. n, le président de l'Assemblée nationale **cubaine**, [CONSIDÉRÉ](#) comme un partisan de l'ouverture, tandis

33. arck un précurseur des nazis. Hitler, **d'ailleurs**, [CONSIDÉRAIT](#) que le chancelier de Guillaume II, mal

34. : que l'Europe de demain cesse tout d'un coup **de** [CONSIDÉRER](#) les pays d'Asie comme un modèle de réf

35. du Rouergue"; les plus modestes se contentent **de** [CONSIDÉRER](#) le lieu comme un écrin de culture et de

36. PIB européen), il paraît, en effet, difficile **de** [CONSIDÉRER](#) que l'UEM puisse être, de ce point de v

37. 'accusant de diviser le peuple et de continuer **de** [CONSIDÉRER](#) comme non israélienne la moitié de la p

CORPORA COLLECTION

considérer

French mixed corpus based on material from 2012
Sentences: 74,823,426 · Types: 7,873,935 · Tokens: 1,468,766,604 more...

Word: **considérer** Number of occurrences: 31,240 Rank: 4,600 Frequency class: 11

See also: [Considérer](#), [considérer](#), [considérer](#), [considérer](#), [Considérer](#), [CONSIDÉRER](#), [considÉrer](#)

Similarity based on Cooccurrences: [penser](#) | [admettre](#) | [envisager](#) | [comprendre](#) | [imaginer](#)

Examples: A défaut d'une meilleure estimation, l'Autorité a...

Cooccurrences:

comme (55,123), peut (18,457), que (12,163), faut (5,844), On (5,707), on (5,519), ne (3,181), l'on (2,655), à (2,415), un (2,356), qu'il (2,276), pas (2,038), une (1,965), il (1,681), de (1,555), donc (1,520), les (1,505), tendance (1,320), est (1,260), ce (1,196), pouvons (1,097), qu'on (1,086), non (992), étant (964), il (952), nous (939), ces (910), doit (891), consiste (716), serait (707), devons (699), le (699), plutôt (656), si (652), pourrait (633), cette (622), façon (607), (598), (592), la (570), point (570), refuse (566), ou (552), entière (536), s'agit (530), manière (522), mais (518), s'accordent (514), convient (510), qu'un (502), qu'une (485), n'est (479), peut-on (451), cela (443), faudrait (441), amène (440), même (440), cas (429), suffit (415), c'est (413)

Approche directe : 1^{re} réflexion sur l'outil

Contextualisation

Comparer les sorties entre sites :
Leur présentation ? Quel type
d'information ?...

- Leur contenu ?
- Leur taille ?

- Textes différents donc langue différente
- Taille différente donc résultats à harmoniser
- Année des textes...

Approche directe : prise en main

Comparer les deux corpus suivants.
Quelles sont vos conclusions ?

Lexicoscope
- Exploration des profils combinatoires -

Sélection du Corpus Requête Paramètres Sessions sauvegardées Guide

Corpus monolingue
Corpus parallèle

Langue fr

- Corpus littéraire du projet phraséotexte - Policier
- Corpus littéraire du projet phraséotexte - Sentimental
- Corpus littéraire du projet phraséotexte - Historique
- Corpus littéraire du projet phraséotexte - Science fiction
- Corpus journalistique français**
- Corpus littéraire français
- Corpus journalistique français (emoBase - étiquettes Connexor)
- Corpus littéraire français (emoBase - étiquettes Connexor)
- Corpus journalistique français (emoBase - étiquettes Connexor - 20M)
- Corpus journalistique français (emoBase - étiquettes Connexor - 16M)

Description

Corpus sélectionné : nombre de mots = **112 280 979** (**299 125** textes)

ScienQuest – Corpus «Écrits scientifiques en français»

ScienQuest

Corpus Textes Recherche Résultats

Scientext

TALN

Actes de TALN

Le corpus [TALN Archives](#) a été collecté en 2013 par Florian Bourdin à partir des différents sites Web des conférences TALN et RÉCITAL (1997-2014). Il s'agit d'un corpus de textes au format pdf, accompagnés de méta-données (notice bibtext et résumé). Un sous-ensemble de 586 articles a ensuite été sélectionné et traité par Ludovic Tanguy, afin d'en extraire le texte intégral, et de l'analyser avec TALISMANÉ. Le [corpus arboré](#) ainsi obtenu contient 2,3 millions de tokens, annotés en parties du discours, en lemmes et en dépendances syntaxiques.

Ce corpus contient *586 textes (2 335 943 mots)*.

Textes d'évaluation

Corpus Scientext - Évaluations du colloque CÉDIL 2010

Ce corpus contient 520 commentaires évaluatifs de relecteurs pour un colloque de jeunes chercheurs en sciences du langage (Colloque international des Étudiants chercheurs en Didactique des Langues et en Linguistique, 2010).

Version 1.0 du corpus, constitué au LIDILEM par Françoise Boch et Achille Falaise, dans le cadre du projet ANR Scientext.

Annotation avec l'analyseur Syntex développé par Didier Bourigault.

Ce corpus contient *570 textes (34 805 mots)*.

Anglais langue étrangère

Corpus Scientext - Écrits en anglais langue étrangère

Ce corpus comporte des travaux d'apprenants universitaires français écrivant en anglais, principalement des étudiants de 2e et 3e année du cursus d'anglicistes apprenant à rédiger de textes argumentatifs longs (4500 mots) qui s'appuient sur des recherches documentaires approfondies.

Version 1.0 du corpus, constitué au LLS par John Osborne, Alice Henderson et Robert Barr, dans le cadre du projet ANR Scientext.

Annotation avec l'analyseur Syntex développé par Didier Bourigault.

Ce corpus contient *272 textes (1 020 146 mots)*.

Écrits scientifiques en anglais

Corpus Scientext - Écrits scientifiques en anglais

Ce corpus a été élaboré par l'équipe LiCorn de l'Université de Bretagne Sud (Geoffrey Williams, Chrystel Millon). Les textes proviennent de la maison d'édition indépendante BioMed Central et portent exclusivement sur la biologie et la médecine.

Annotation avec l'analyseur Syntex développé par Didier Bourigault.

Ce corpus contient *7 564 textes (35 244 378 mots)*.

S'interrogent sur leurs besoins

- Questions sémantiques :
onomasiologiques
 - Comment écrire mes questions de recherche?
 - Comment citer un auteur ?
 - ...

Question sémantique sur le web

Recherche sémantique

ScienQuest – Corpus «Écrits scientifiques en français»

Langue: [fr] [en]

UNIVERSITÉ Stendhal

Corpus | Textes | Recherche | Résultats | Historique | Connexion | À propos | Aide

Sémantique | Libre | Avancée

- Auteurs cités**
 - Citations
- Autour des hypothèses**
 - Formulation d'une hypothèse
 - Validation d'une hypothèse
- Dénomination**
 - Verbe + "sous le nom"
 - Définir comme
 - Donner le nom
 - Entendre par
 - Nommer
- Propositions propres de l'auteur**
 - Verbes de choix et d'intention
 - Verbes de résultats et apports scientifiques
- Évaluation et opinion**
 - Adjectifs d'opinion
 - Adjectifs d'évaluation
 - Adverbiaux d'opinion
 - Verbe modal d'opinion
 - Noms d'opinion
 - Verbes d'opinion

Vers l'écriture

Répondez aux questions suivantes

1

1. D'après vous, cet extrait vient de quelle partie de l'article ? Pourquoi ?
A. résumé // B. introduction // C. théorie // D. méthodologie // E. conclusion
2. Découpez l'extrait en paragraphes et justifiez votre choix en précisant les mots et les thématiques de votre découpage à l'aide du tableau suivant :

Contexte	Sujet	Objectif	Question

3. Trouvez des expressions synonymes pour dire « on tente de... » à l'aide de l'outil CNTRL en ligne : <http://www.cnrtl.fr/synonymie/>

Précisions ensuite sur introduction / questions de recherche / citer un auteur...

ométrie

2

Découpage

Contexte

L'analyse linguistique que nous proposons s'inscrit en réponse à une demande de la SNCF relative à la perception du confort global en train. Les objectifs du projet sont d'identifier les propriétés sémantiques du confort en train et leurs relations de dépendance, à partir d'analyses linguistiques et cognitives.

Sujet

- **Cet article porte sur** la manière dont les formes linguistiques en contexte, utilisant les ressources de la langue mises en discours, renseignent sur les structures cognitives construites à partir des perceptions sensorielles.

Objectif

- **On tente ici d'identifier**, à partir de l'analyse des formes adjectivales, les représentations individuelles et partagées qui se construisent dans les discours des voyageurs, lorsqu'on les interroge sur leur expérience sensible du confort.

Question

- **La première question que l'on se pose alors est la suivante** : le confort en train est-il une catégorie cognitive de ce type ?

46

Liste des éléments

3

- Présenter le sujet : → **Cet article porte sur**
- Introduire les questions pour cet article : → **La première question que l'on se pose alors est la suivante**
- Présenter le contexte : → **L'analyse linguistique que nous proposons s'inscrit (en / dans)**
- Présenter les objectifs de l'article : → **On tente ici de**

Trouvez-en d'autres

Démarche FOS

Étudiants de Master pour rédiger leurs mémoires

1/ Commande → analyse de la demande

Manque d'éléments spécifiques aux écrits académiques

2/ Analyse des besoins / premières hypothèses

3/ Contact avec les acteurs du terrain / collecte des données / confirmation ou infirmation des hypothèses

Attentes universitaires
Création de corpus des genres d'écrits universitaires

4/ Traitement des données / analyse des discours

Analyse de la phraséologie transdisciplinaire

5/ Élaboration du programme de formation

Programme de formation à la méthodologie de la recherche

6/ Élaboration des activités didactiques

Utiliser les corpus comme documents authentiques

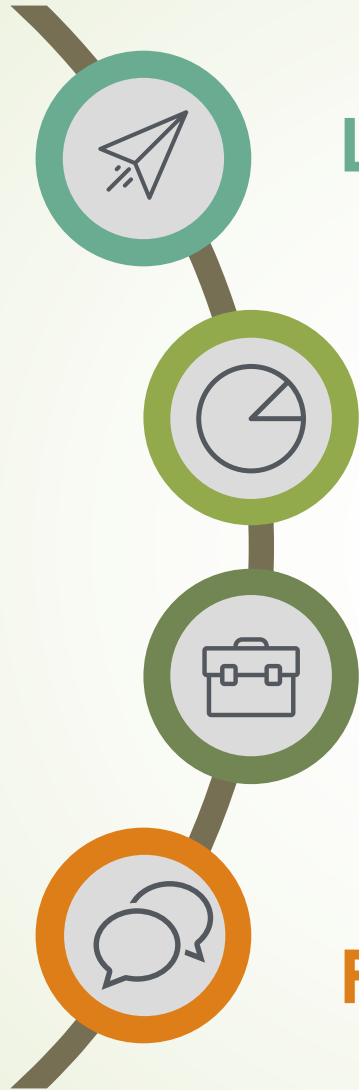
Cavalla - 2021 - BCL, séminaire Logométrie

Récapitulatif

Dispositif pédagogique :

1. Entrer par le « déjà là » : recherche en ligne
2. Présenter des outils d'aide à la rédaction
3. Conduire vers la langue spécialisée

PRÉSENTATION



LES CORPUS EN CLASSE DE LANGUE

CORPUS ET ÉCRITS SCIENTIFIQUES

CORPUS ET ÉMOTIONS

PERSPECTIVES

1. Repérage du lexique : corpus numériques

Quel lexique repérer ?

3 corpus : 3 discours

ScienQuest – Corpus *Écrits scientifiques en français*

Lexicoscope
- *Exploration des profils combinatoires* -

Frantext intégral

Base de données intégrale Frantext

- Textes **SCIENTIFIQUES**
- 2000-2013
- 4,8 millions de mots

- Textes **JOURNALISTIQUES**
- 2007-2008
- 109 millions de mots

- Textes **LITTÉRAIRES**
- 2000-2014
- 11 millions de mots

Lexies choisies

amitié	étonnement
amour	inquiétude
angoisse	joie
bonheur	peur
colère	plaisir

Fréquences

	Scientext		Frantext		Emoconc	
	4800000 mots		11000000 mots		109000000 mots	
	Nb occ.	‰	Nb occ.	‰	Nb occ.	‰
amitié	37	0,008	669	0,06	9494	0,09
amour	199	0,04	4343	0,4	29439	0,3
angoisse	100	0,02	895	0,08	5822	0,05
bonheur	39	0,008	1407	0,1	13767	0,1
colère	131	0,03	776	0,07	14232	0,1
étonnement	12	0,002	271	0,02	1306	0,01
inquiétude	28	0,006	460	0,04	15461	0,1
joie	47	0,01	1189	0,1	8500	0,08
peur	288	0,06	2874	0,2	21307	0,2
plaisir	134	0,03	2256	0,2	18438	0,17

2. Classement du lexique : sémantique

Quel lexique ?

Les émotions dans les situations de communication

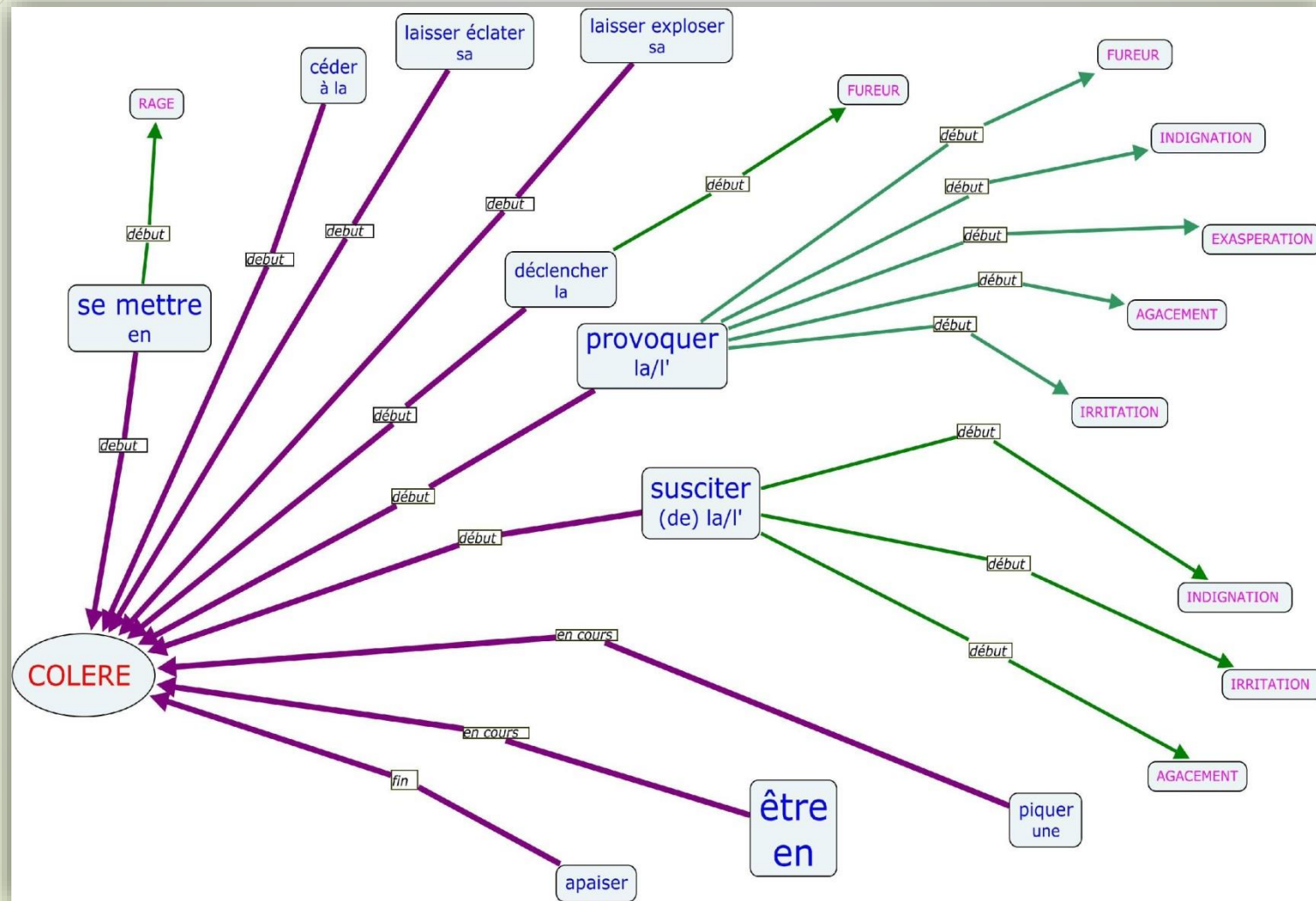


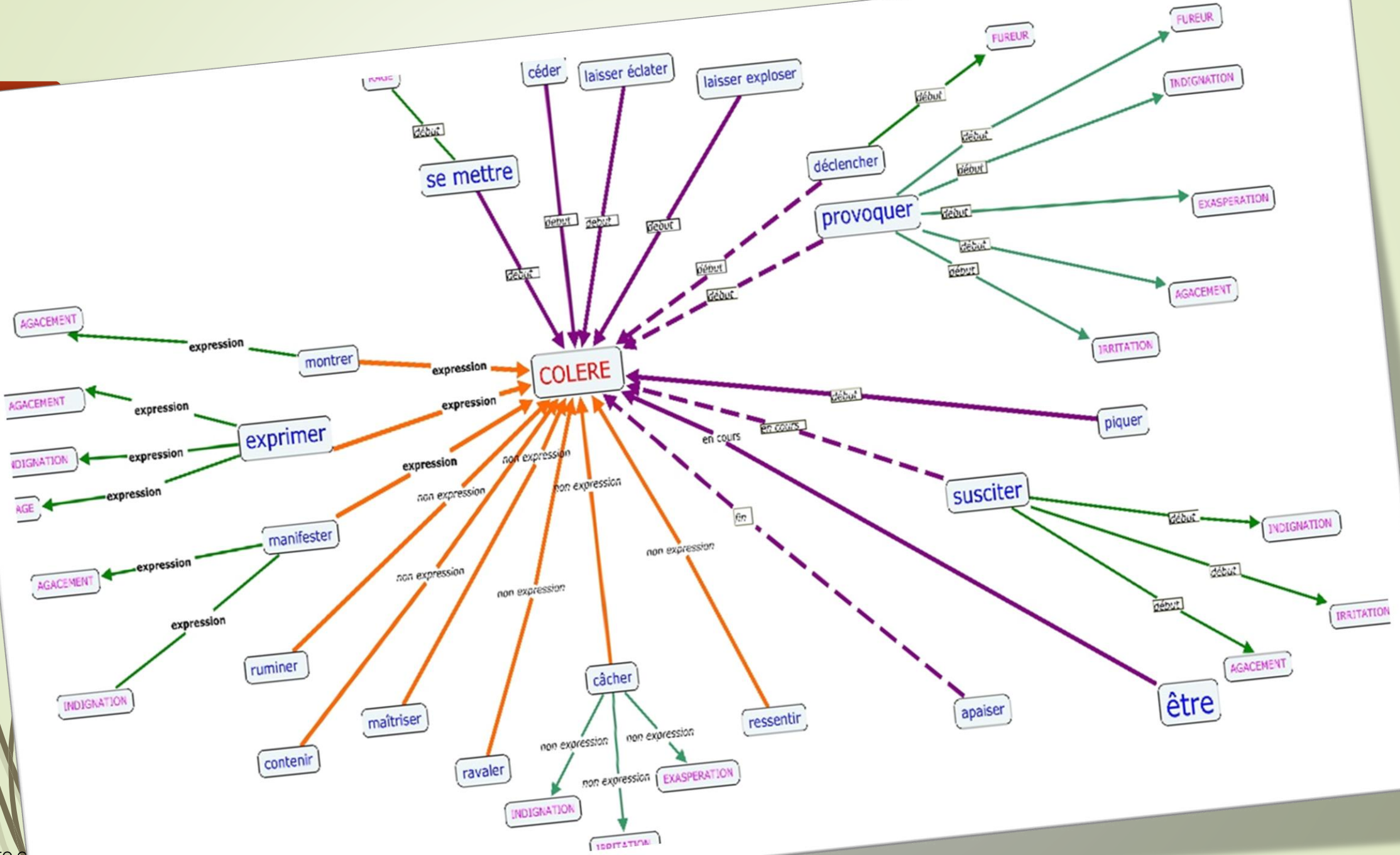
3 objets à enseigner

- Savoir **décoder** les émotions des interlocuteurs
- Savoir **exprimer** des émotions

- Mise en situation
- Lexique
- Gestuelle

forme et sens





Rassembler pour avancer



- contexte, réseau sémantico-culturel (mémorisation)
- forme, sens, fréquence (Repérage)
- sens, contexte (Classement)

Classement des éléments émotionnels

Ulrich poussa un cri d'appel aigu, vibrant, prolongé. La voix s'envola dans le silence de mort où dormaient les montagnes ; elle courut au loin, sur les vagues immobiles et profondes d'écume glaciale, comme un cri d'oiseau sur les vagues de la mer ; puis elle s'éteignit et rien ne lui répondit. (Maupassant, 1887: *l'auberge dans le horla*)

Émotions	UL simples	Formes figées	Éléments associés	Représentations
Peur Angoisse Détresse	Cri	Pousser un cri Silence de mort Cri d'appel	Cri d'appel aigu / vibrant / prolongé Écume glaciale	Cri : peur ? Dormaient les montagnes, les vagues immobiles : pétrifié/paralysé de peur ? Vagues profondes : peur de ce qu'on ne voit pas ? Écume glaciale : froid désagréable, paralysant ?

Recherches et comparaisons dans les corpus

Lexicoscope

- Exploration des profils combinatoires -

- Exploration des profils combinatoires -

Frantext intégral

Base de données intégrale Frantext

Contexte des émotions

Corpus

- Peu de lexique
- Contexte à voir

Constat

- Difficile à décoder
- Difficile à encoder

Conséquence

- Définition
Cadre
Emotionnel



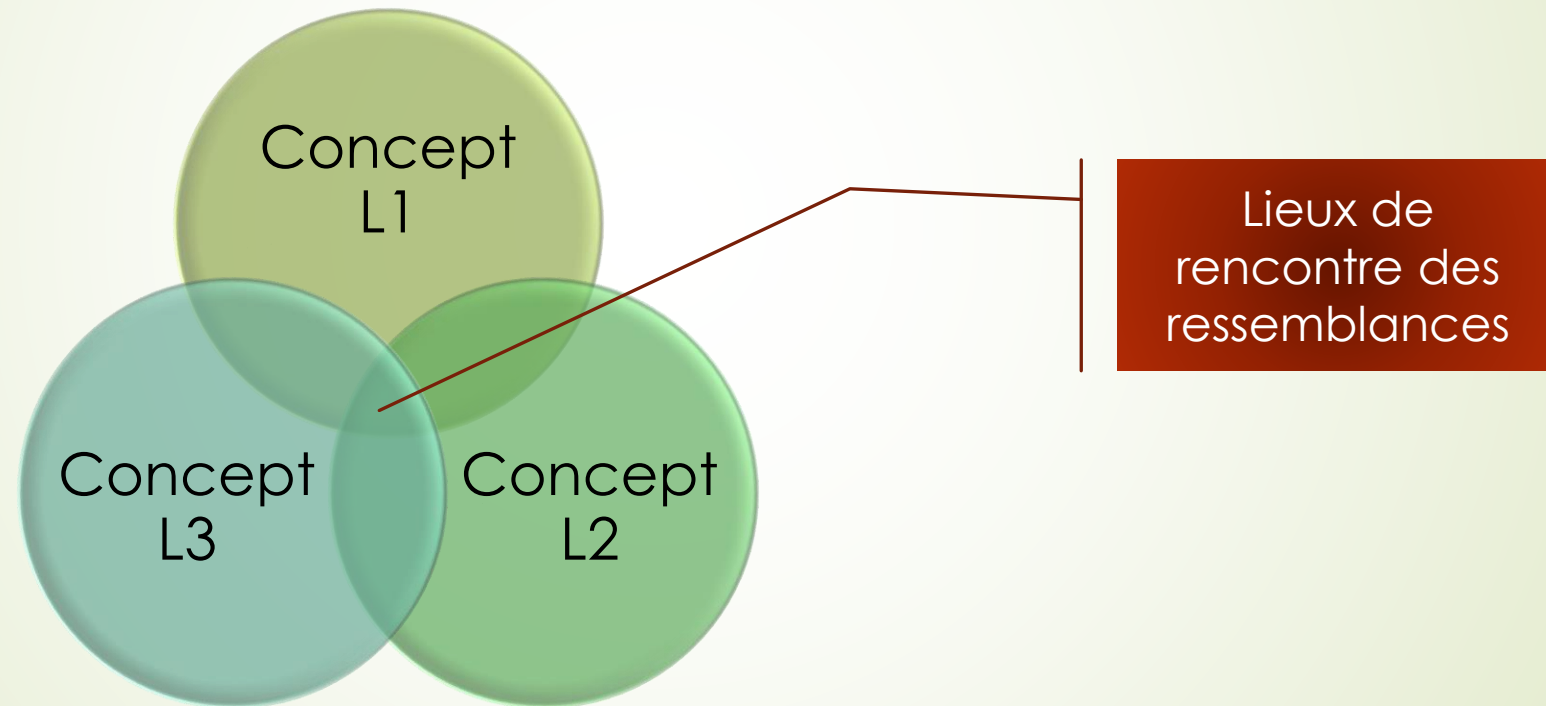
Image par Gordon Johnson de Pixabay

4. Cadre émotionnel

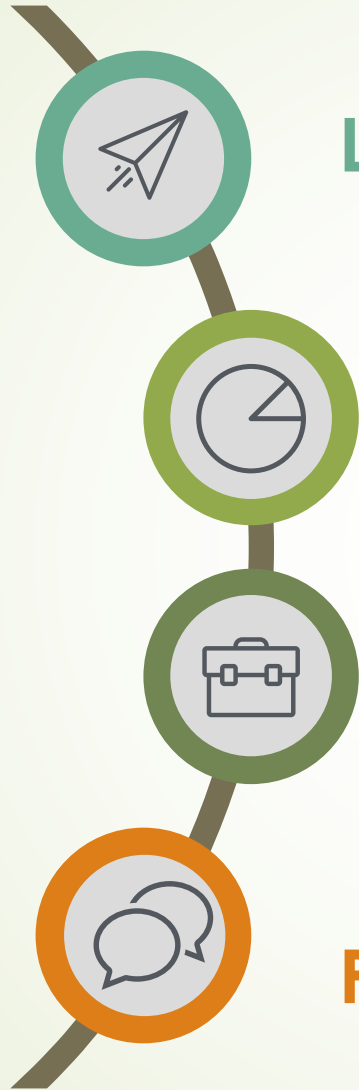
Le cadre émotionnel : définition

Nous utilisons le terme de cadre émotionnel pour rappeler le fait que, dans le récit fictionnel, les lieux décrits sont fréquemment investis, « psychologisés » et jouent d'une certaine manière le rôle d'actant, au sens de la sémiotique narrative. (Grossmann, Boch, Cavalla, 2008 : 199)

Comparer les concepts



PRÉSENTATION



LES CORPUS EN CLASSE DE LANGUE

CORPUS ET ÉCRITS SCIENTIFIQUES

CORPUS ET ÉMOTIONS

PERSPECTIVES

Développer des corpus



Corpus scientifiques par disciplines



Corpus avec recherche sémantique

Université Côte d'Azur
UMR 7320 : Bases, Corpus, Langage
Séminaire
Logométrie. Corpus, Traitements, Modèles

Merci

Corpus numériques en FLE

Cristelle CAVALLA

cristelle.cavalla@sorbonne-nouvelle.fr

6 mai 2021

**Sorbonne
Nouvelle**  DILTEC - EA 2288
Didactique des langues,
des textes et des cultures