

Recherche Zen

Séance 1 : Questions et hypothèses de recherche

Carlos Ramisch and Manon Scholivet

Partly based on the course by Adeline Paiement

4 octobre 2023

Qui sommes nous ?

Manon Scholivet

Qui sommes nous ?

Manon Scholivet

- Thèse soutenue en 2021

Manon Scholivet

- Thèse soutenue en 2021
- Passionnée de science propre et carrée

Manon Scholivet

- Thèse soutenue en 2021
- Passionnée de science propre et carrée
- Méthodologie de la recherche apprise sur le tas

Qui sommes nous ?

Manon Scholivet

- Thèse soutenue en 2021
- Passionnée de science propre et carrée
- Méthodologie de la recherche apprise sur le tas

Carlos Ramisch

Qui sommes nous ?

Manon Scholivet

- Thèse soutenue en 2021
- Passionnée de science propre et carrée
- Méthodologie de la recherche apprise sur le tas

Carlos Ramisch

- Thèse soutenue en 2012
- Fan d'expressions idiomatiques

Qui sommes nous ?

Manon Scholivet

- Thèse soutenue en 2021
- Passionnée de science propre et carrée
- Méthodologie de la recherche apprise sur le tas

Carlos Ramisch

- Thèse soutenue en 2012
- Fan d'expressions idiomatiques
- Méthodologie de la recherche apprise sur le tas

- Séance 1 - Questions et hypothèses de recherche
- Séance 2 - Contextualisation bibliographique
- Séance 3 - Expériences
- Séance 4 - Analyses
- Séance 5 - Communication
- Séance 6 - Pratiques éthiques

Supports :

<https://pageperso.lis-lab.fr/carlos.ramisch/?page=recherchezen>

La Science : c'est quoi ?

Définition de la science pour vous ?

La Science : c'est quoi ?

Sont scientifiques les énoncés qui se révèlent certains

La Science : c'est quoi ?

Sont scientifiques les énoncés qui se révèlent certains

→ On a été sûr.e.s que le soleil tourne autour de la Terre. Et pourtant ...

La Science : c'est quoi ?

Sont scientifiques les énoncés qui se révèlent certains

→ On a été sûr.e.s que le soleil tourne autour de la Terre. Et pourtant ...

- Ne sont scientifiques que les domaines pouvant se remettre en question.
- Tous les modèles sont faux, mais certains sont utiles
- Les modèles actuels seront remplacés par de meilleurs modèles demain

Science = méthode scientifique

Science = méthode scientifique

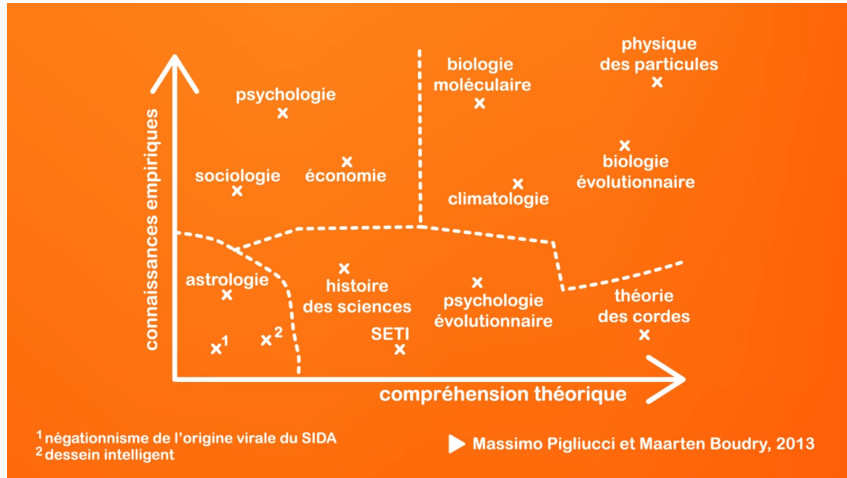
- Les méthodes scientifiques sont aussi utilisées par les journalistes, et certains chasseurs de fantômes..
- Les méthodes évoluent... Ex : Utilisation de groupes placebo en médecine

Émettre des hypothèses testables

Émettre des hypothèses testables

→ L'astrologie est donc de la science, mais pas la théorie des cordes

- Aujourd'hui : vision moins binaire de la science. La "scientificité" d'un domaine se situe sur un spectre : des disciplines considérées scientifiques, d'autres plutôt pseudo scientifiques.
- Certains critères bien que pas parfaits, restent une intuition forte
 - Testabilité
 - Remise en question du domaine, ou croyances dogmatiques ?



Science & pseudoscience : les différencier, pas gagné ! du Chat Sceptique

Bravo : vous venez de faire un peu d'Épistémologie !

C'est la science de la connaissance.

Organisation de la recherche scientifiques :

- De nombreux rituels :
 - Articles scientifiques
 - Présentations en conférences
 - Soutenances (thèses, HDR, ...)
 - Relectures par les pairs
 - ...
- Des institutions :
 - Universités
 - CNRS
 - Entreprises et laboratoires privés
 - Sociétés savantes
 - ...

Pourquoi ? vs. comment ?

Pourquoi ? vs. comment ?

→ Ou plutôt "comprendre" vs. "résoudre un problème"

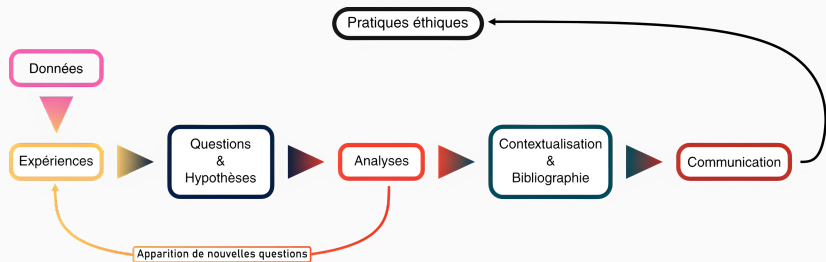
Pourquoi ? vs. comment ?

→ Ou plutôt "comprendre" vs. "résoudre un problème"

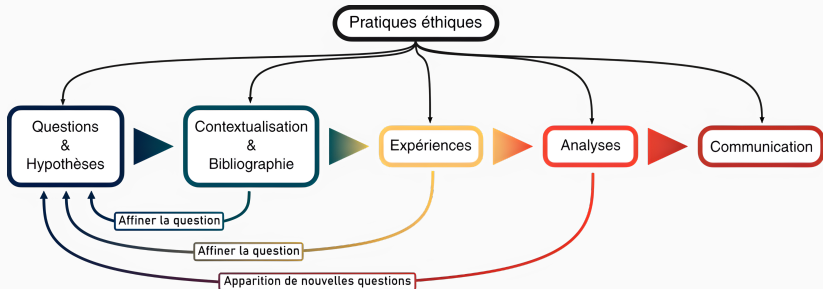
Fil directeur du cours : ne pas hiérarchiser

- La recherche supérieure à d'autres formes de pensée → **Non**
- La "bonne recherche" (la nôtre, forcément) meilleure que celle des autres (les autres font n'importe quoi) → **Non**

La méthode : en pratique



La méthode : dans l'idéal



Début de recherche, généralement chaotique...

- A l'entame d'une recherche, le chercheur est souvent perdu :
 - vague idée sur l'objet d'étude
 - ... mais sans savoir comment s'y prendre
 - ... et en ayant la crainte de mal engager le travail

⇒ Chaos originel

- Une situation éprouvante, mais inévitable :
 - La recherche c'est :
 - ⇒ un cheminement vers une meilleure connaissance
 - ⇒ quelque chose qui par définition "se cherche" donc comprend hésitations, errements, incertitudes

Dangers du chaos originel 1/3

Le chaos originel, attention à :

Dangers du chaos originel 1/3

Le chaos originel, attention à :

- L'impasse aux hypothèses

- La "hype"

- ⇒ Fuite en avant encouragée par la croyance que l'usage de techniques savantes détermine la valeur intellectuelle du travail
- ⇒ Effet de mode : tester les dernières avancées technologiques
- ⇒ "Je vais appliquer tous les modèles de machine learning disponibles dans cette bibliothèque, puis on verra"



Dangers du chaos originel 2/3

Le chaos originel, attention à :

- L'impasse aux hypothèses
 - Le tsunami de données
 - ⇒ Collecte des données avant formulation des hypothèses
 - ⇒ Question sociétale - toujours plus, plus vite, des données partout !
 - ⇒ "Voici mes données, faites-moi du machine learning dessus"



Dangers du chaos originel 3/3

Le chaos originel, attention à :

- **Too many ideas, not enough me's**
 - ⇒ Ambition démesurée + confusion la plus complète
 - ⇒ "Je vais créer un traducteur parfait qui fonctionne dans toutes les langues"



Question de recherche : pour quoi faire ?

La question de recherche permet de surmonter le "chaos originel"

- **Fonction :**

- Clarification des intentions
- Première étape de la rupture avec les préjugés

- **Modalités :**

- Énoncer le projet de recherche sous la forme d'une question
- Par cette question on tente d'exprimer ce qu'on cherche à savoir ou à mieux comprendre

⇒ Premier **fil conducteur** de la recherche

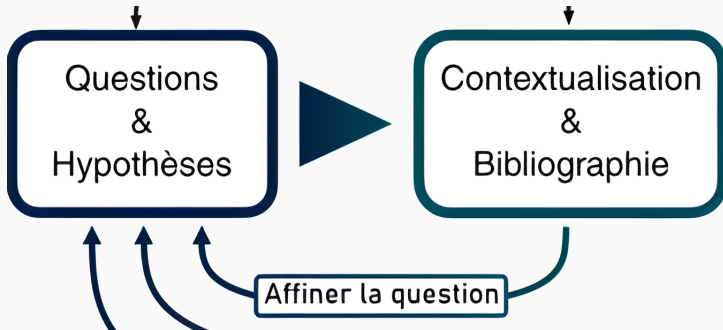
- Problème :
 - Écart entre une **situation actuelle insatisfaisante** et une **situation visée**
 - La recherche entreprise doit permettre de **réduire cet écart** en répondant à une question de recherche
- Problématiser :
 - Cheminer d'une **thématique générale** à une **question de recherche**
 - S'appuie sur les données de la **revue de la littérature** et les expériences et intuitions personnelles
 - Processus itératif

Analyser : découper la montagne

Transformer un problème énorme et vague en une série de problèmes plus simples et plus précis



L'oeuf et la poule



Problématiser : comment, dans quel contexte ?

Problème

Ex : Les systèmes de traduction automatique sont utiles mais toujours pas parfaits. Plusieurs pistes d'amélioration s'offrent à nous.

Question de recherche

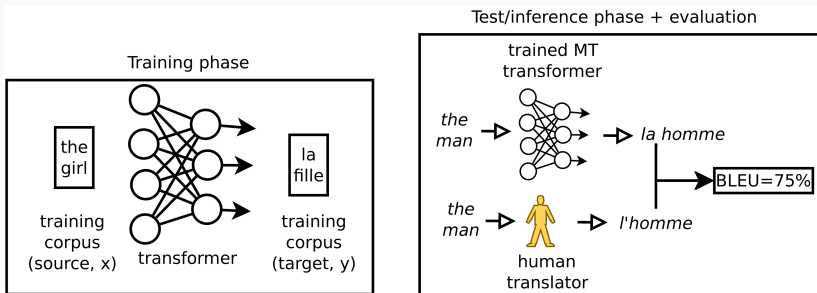
Ex : Est-ce que la qualité de la traduction automatique (mesurée par le score BLEU sur un jeu de données standard anglais-français) augmente lorsqu'on dispose de plus de données pour apprendre un système de traduction basé sur les Transformers ?

Revue de littérature

Identifie un déficit de connaissances en lien avec le problème

Ex : Les articles en traduction automatique utilisent tous les données d'entraînement disponibles sans se poser la question de l'impact de la quantité de données.

La traduction automatique en 30 secondes



Question de recherche

Question de recherche

Est-ce que la qualité de la traduction automatique (mesurée par le score BLEU sur un jeu de données standard anglais-français) augmente lorsqu'on dispose de plus de données pour apprendre un système de traduction basé sur les Transformers ?

Des objectifs

- Entraîner un système français-anglais sur 10k phrases
- Entraîner un système français-anglais sur 20k phrases
- ...
- Comparer le score BLEU des systèmes, observer la tendance

Des hypothèses

- Le score BLEU sera plus élevé quand on utilise plus de données d'entraînement
- L'augmentation du score BLEU n'est pas linéairement proportionnelle à la quantité de données ajoutées

Question de recherche

Question de recherche

principale

Est-ce que la qualité de la traduction automatique (mesurée par le score BLEU sur un jeu de données standard anglais-français) augmente lorsqu'on dispose de plus de données pour apprendre un système de traduction basé sur les Transformers ?

Des objectifs

- Entraîner un système français-anglais sur 10k phrases
- Entraîner un système français-anglais sur 20k phrases
- ...
- Comparer le score BLEU des systèmes, observer la tendance

Des hypothèses

- Le score BLEU sera plus élevé quand on utilise plus de données d'entraînement
Vrai/faux ? question secondaire
- L'augmentation du score BLEU n'est pas linéairement proportionnelle à la quantité de données ajoutées
Vrai/faux ? question secondaire

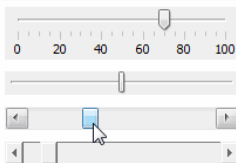
Question secondaire / sous-question : plus courte et moins complexe

Réponses à la question de recherche

- Les questions secondaires définissent presque complètement la ou les expériences à faire
- Le plus souvent : pas de réponse ferme à la question principale
- On apporte des réponses aux questions secondaires, petit à petit on a des éléments de réponse pour la question principale

Curseur de croyances

- Les réponses aux questions secondaires ajustent le curseur de croyances de la question principale
- Le curseur de la question principale n'atteindra peut-être jamais 0 ou 1.



Qu'est-ce qui fait une BONNE question de recherche ?

Passer d'une idée de recherche à une question de recherche :

Idée	Question
Créative	Méthodique
Impulsive	Réfléchie
Palpitante	Avec retenue
Vague	Formalisée ou conceptualisée
Non vérifiée	Vérifiable
Non validée	Validée
Illimitée	Limite de temps, de cadre et de budget

(Martinez and Berkhout, 2009)

Qu'est-ce qui fait une **BONNE** question de recherche ?

Qu'est-ce qui fait une BONNE question de recherche ?

- Pertinente

Qu'est-ce qui fait une BONNE question de recherche ?

- Pertinente
- Ciblée et Claire

Qu'est-ce qui fait une BONNE question de recherche ?

- Pertinente
- Ciblée et Claire
- Faisable / Aboutissant à un protocole réalisable

- Ai-je bien défini le champ de mon étude ?
- Ai-je une bonne connaissance du champ de mon étude et de la littérature qui s'y rapporte ?
- Quelles sont les questions de recherche importantes qui se posent dans ce champ ?
- Quels sont les secteurs du champ d'étude qui méritent une exploration complémentaire ?
- Mon étude comblera-t-elle le manque d'information ?
Permettra-t-elle une meilleure compréhension du champ d'étude ?
- N'y a-t-il pas déjà de nombreuses études qui ont été menées dans ce domaine ?

- Mon étude n'a-t-elle pas déjà été réalisée ? Si oui, apportera-t-elle des informations manquantes ou une confirmation d'informations incertaines ?
- Est-ce le bon moment pour répondre à cette question ? N'est-ce pas une question trop brûlante ou, au contraire, qui sera rapidement obsolète ?
- Mon étude permettra-t-elle un progrès dans mon champ d'expertise ? Aura-t-elle un impact sur la société ?

⇒ Question de l'**impact** de la recherche

Une question claire est :

- Ciblée
- Précise : son sens ne prête pas à confusion
 - Les termes utilisés sont définis au préalable
 - La formulation est la plus limpide possible
- Contre-exemple : **Quel est l'impact des améliorations récentes de l'IA sur la recherche en traitement automatique des langues (TAL) ?**
- Quelles améliorations ? Quelle(s) recherche(s) ? Quelle(s) branche(s) du TAL ?

Vérifier la compréhension de plusieurs personnes \Rightarrow identiques ?
divergentes ?

Une question claire est :

- Concise
- Univoque : une seule interprétation possible
 - Les termes utilisés sont définis au préalable
 - La formulation est la plus limpide possible

Contre-exemple : Dans quelle mesure l'augmentation des pertes d'emplois dans le secteur du transport aérien explique-t-elle le maintien de participation publique dans les compagnies aériennes destinée non seulement à soutenir ce secteur mais aussi à diminuer les risques de conflits sociaux que cette situation porte en elle ?

Question trop longue, embrouillée, comporte des suppositions

Dédoublage de la question sur la fin : que cherche t'on en priorité ?

Très souvent : question initiale trop vaste / Univoque : une seule interprétation possible

Exemple : Pourquoi la qualité des traducteurs automatiques varie autant selon les langues ?

Question très vaste !

Beaucoup d'angles d'approche différents \Rightarrow méthodologie difficile à définir

\Rightarrow Éviter de débiter une question de recherche principale par « pourquoi »

Autre problème : « autant » est subjectif et peu précis, comme les termes « en général », « fréquent », « beaucoup », « peu », ...

Très souvent : question initiale trop vaste / vague

Autre exemple :

Avoir plus de données disponibles permet il d'améliorer la performance des systèmes de traduction automatique ?

Très souvent : question initiale trop vaste / vague

Autre exemple :

Avoir plus de données disponibles permet il d'améliorer la performance des systèmes de traduction automatique ?

⇒ C'est quoi « améliorer » ? Quel est le but exact ?

Différentes métriques mettent l'accent sur différents aspects des résultats : precision, recall, accuracy, AUC, F1 score...

Très souvent : question initiale trop vaste / vague

Autre exemple :

Avoir plus de données disponibles permet il d'améliorer la performance des systèmes de traduction automatique ?

⇒ Quel(s) système(s) de traduction automatique ? Transformer ? Statistique ? Fondé sur des règles ? etc.

Différents systèmes peuvent être plus ou moins dépendants de la quantité de données disponibles.

Très souvent : question initiale trop vaste / vague

Autre exemple :

Avoir plus de données disponibles permet il d'améliorer la performance des systèmes de traduction automatique ?

⇒ À quel point les résultats de cette étude pourront être généralisés à d'autres contextes ?

Les résultats vont dépendre de la qualité des données.

Une version plus ciblée :

Est-ce que la qualité de la traduction automatique (mesurée par le score BLEU sur un jeu de données standard anglais-français) augmente lorsqu'on dispose de plus de données pour apprendre un système de traduction basé sur les Transformers ?

- Objectif : étudier la corrélation entre quantité de données et qualité de traduction pour améliorer la précision de la traduction
- Schéma de l'étude : prédiction par un transformer sur des données anglais-français
- Validité : jeux de données standard anglais-français

Question trop vaste / vague :

⇒ Doit être réduite

⇒ Processus commun et frustrant pour les jeunes chercheurs

1. Revue de littérature → identifier les inconnues en attente de réponse
2. Formuler plusieurs questions
3. Choisir la plus pertinente et la mieux ciblée

Question trop vaste / vague :

⇒ Doit être réduite

⇒ Processus commun et frustrant pour les jeunes chercheurs

1. Revue de littérature → identifier les inconnues en attente de réponse
2. Formuler plusieurs questions
3. Choisir la plus pertinente et la mieux ciblée

Attention aux questions trop ciblées et restrictives : impact restreint

Ai-je les moyens de répondre à ma question de recherche ?

- Collecte des données (disponibilité des sources, matériel...)
- Expertise technique
- Temps
- Financement
- Équipe et gestion d'équipe

Contraintes matérielles et de temps souvent sous-estimées !

Puis-je trouver des sources de financement ?

Y a-t-il des organismes intéressés par mon travail qui pourraient le financer ?

Mon étude aura-t-elle un impact dans son champ et dans la société ?

Puis-je convaincre les organismes de financement que je suis en mesure de mener à bien la recherche ?

- Expertise
- Résultats préliminaires
- Historique de projets menés à bien

Une proposition de structuration du travail

- **Q**uestion de recherche – multi-niveaux, sous-questions précises
- **H**ypothèses – associées à chaque sous-question
 - Si j’obtiens tel résultat, ma conclusion sera X, sinon Y
- **J**ustification – mes questions de recherche sont
 - pertinentes,
 - potentiellement utiles,
 - pas encore couverte dans la littérature
- **E**xpérience – associée à chaque sous-question
 - Données
 - Conditions expérimentales
 - Métriques d’évaluation
 - ...

Exemple : poulpes et canicule

Question de recherche : Quels sont les impacts des canicules marines sur les habitudes alimentaires du poulpe commun ?

Q1 : La quantité de nourriture ingérée par les poulpes pendant les périodes de canicule marine diminue-t-elle ?

- Hypothèse **H1** : Oui
 - H1 est validée si en moyenne, les poulpes mangent significativement moins lors des canicules.
- Justification **J1** : Les humains mangent moins quand il fait chaud
- Expérience **E1** : comparer les quantités moyennes pendant canicule vs. sans canicule (1 semaine avant)

Exemple : poulpes et canicule

Question de recherche : Quels sont les impacts des canicules marines sur les habitudes alimentaires du poulpe commun ?

Q2 : La quantité de nourriture ingérée chaque année selon la période (été vs. hiver) changent lors des années caniculaires ?

- Hypothèse H2 : Oui
 - H2 validée si la quantité moyenne de nourriture consommée par les poulpes en hiver est significativement plus importante les années caniculaires
- Justification J2 : Les poulpes compensent les périodes de "famines caniculaires" en mangeant plus l'hiver
- Expérience E2 : on compare la quantité moyenne en été vs. hiver les années de canicules marines et les années classiques

- Le niveau de précision (granularité) d'une question de recherche dépend du contexte du travail
 - Thèse : question de recherche globale un peu générale, avec sous-questions très précises
 - Article : question de recherche précise pour une contribution ciblée

- Cours d'Adeline Paiement
- Chaîne Youtube Le Chat Sceptique
- Chaîne Youtube Hygiène Mentale
- Chaîne Youtube Monsieur Phi
- Google images
- Midjourney