

Recherche Zen

Séance 2 : Contextualisation et bibliographie

Carlos Ramisch and Manon Scholivet

Partly based on the course by Adeline Paiement

11 octobre 2023

Introduction

Chercher des ressources

Analyser les ressources trouvées

J'ai ma question de recherche ! ... Et maintenant ?

Il est temps de s'intéresser aux travaux des pairs.
C'est l'heure de faire :

- Une **recherche bibliographique** !
- Une **revue de la littérature** !

Oui c'est la même chose... Non ?

Quelle est la différence ?

- **Recherche bibliographique** :
 - **Acquisition, approfondissement** des connaissances sur un sujet précis
- **Revue de la littérature (survey)** :
 - **Synthèse de l'état de l'art** et des connaissances dans le domaine

Quelle est la différence ?



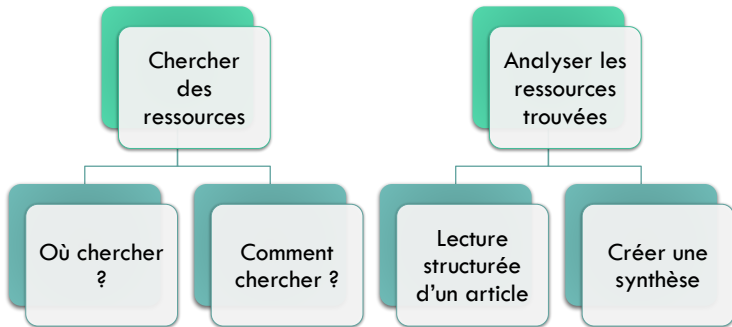
- Une recherche bibliographique n'est PAS une revue de la littérature.
- Mais une revue de la littérature implique de faire une recherche bibliographique.

Quand faire l'un ou l'autre ?

- Recherche bibliographique :
 - Avant tout travail de recherche
 - Pendant l'affinage de la question de recherche
- Revue de la littérature (survey) :
 - La dernière revue est trop ancienne OU
 - Une telle revue n'a jamais été faite ET
 - J'ai le temps de me poser et faire le point

Comment faire l'une ou l'autre ?

- **Recherche bibliographique** :
 - Pas de méthode stricte
 - Idée de point de départ : un article amène à lire un autre article, tel un dictionnaire dont une première définition mène à une seconde, une troisième, etc...
- **Revue de la littérature** :
 - Revue **narrative** : pas de méthode stricte
 - Revue **systématique** : méthodologie existante (p.ex. PRISMA)
 - Recherche en priorité des revues déjà existantes, des méta analyses
 - Définir (à priori ou au fur et à mesure) la portée de la revue

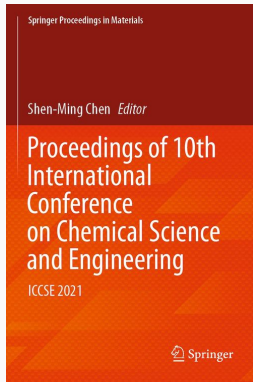


Introduction

Chercher des ressources

Analyser les ressources trouvées

- Journaux et actes de conférences (proceedings)
 - Les BU ont souvent des abonnements aux revues payantes



The image shows the logo for arXiv, a preprint server. The word "arXiv" is written in a bold, black, sans-serif font. A white diagonal slash with a black outline crosses through the 'X' from the top-left to the bottom-right.

Archives

- Nationales : HAL en France
- Internationales : en général gérées par les universités
- Spécialisées : DBLP (informatique), ACL Anthology (TAL)...



- Autres :
 - Pages web des labos, équipes, chercheur.se.s du domaine
 - Parfois on trouve les “preprint” sur des sites perso
 - Twitter, réseaux sociaux (selon les domaines)

- Sci-Hub



Comment chercher des ressources ?

- Google Scholar & SSRN
 - Choix judicieux des mots clé
 - Identifier les articles de référence
 - Nombre de citations et téléchargements

Comment chercher des ressources ?

- Google Scholar & SSRN
 - Choix judicieux des mots clé
 - Identifier les articles de référence
 - Nombre de citations et téléchargements
- Newsletters des journaux, éditeurs
 - IEEE, Elsevier, etc.

Comment chercher des ressources ?

- Google Scholar & SSRN
 - Choix judicieux des mots clé
 - Identifier les articles de référence
 - Nombre de citations et téléchargements
- Newsletters des journaux, éditeurs
 - IEEE, Elsevier, etc.
- Alertes mail sur mot-clé : Google Scholar, ArXiv, etc.

Comment chercher des ressources ?

- Google Scholar & SSRN
 - Choix judicieux des mots clé
 - Identifier les articles de référence
 - Nombre de citations et téléchargements
- Newsletters des journaux, éditeurs
 - IEEE, Elsevier, etc.
- Alertes mail sur mot-clé : Google Scholar, ArXiv, etc.
- Mailing lists thématiques et des sociétés savantes
 - GDR, projets ANR, associations, etc.
 - Annonces d'actes de conférences et numéros de revues

Comment chercher des ressources ?

- Google Scholar & SSRN
 - Choix judicieux des mots clé
 - Identifier les articles de référence
 - Nombre de citations et téléchargements
- Newsletters des journaux, éditeurs
 - IEEE, Elsevier, etc.
- Alertes mail sur mot-clé : Google Scholar, ArXiv, etc.
- Mailing lists thématiques et des sociétés savantes
 - GDR, projets ANR, associations, etc.
 - Annonces d'actes de conférences et numéros de revues
- Suivi de liens
 - Trouvés dans les revues de littérature de travaux antérieurs

Comment chercher des ressources ?

- Connected papers

Search... Expand

Origin paper
Language Models are Few-Shot Learners
Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J... 2020

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee... 2019

Language Models are Unsupervised Multitask Learners
Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, Ilya... 2019

RoBERTa: A Robustly Optimized BERT Pretraining Approach
Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi,... 2019

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova 2019

PaLM: Scaling Language Modeling with Pathways
Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten... 2022

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and...
M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel... 2019

HuggingFace's Transformers: State-of-the-art Natural Language Processing
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond,... 2019

RoBERTa: A ROBUSTLY OPTIMIZED BERT PRE-

Created on Sep 15 2023

2014 2020 2023

Les ressources bibliographiques doivent être :

- Utiles
- Récentes
- Originales

Comment chercher des ressources utiles ?

- **Attention aux dates !**
 - Recherche historique : toutes
 - Recherche des dernières avancées : < ~5 ou 10 ans
- Garder en tête qu'on ne pourra **pas tout lire !**
 - Priorité par pertinence et impact
 - Ordre de grandeur : ~50-100 articles pour un doctorat
 - On s'arrête quand les nouveaux articles semblent prévisibles
 - connaissance suffisante du domaine

Comment chercher des ressources récentes ?

- Ce qui compte comme récent dépend du domaine
 - Psychologie cognitive : ~10 dernières années
 - Machine learning : ~3 derniers mois

Exemple : ACL reviewing policies

If you are aware of relevant publicly available research that has not been cited [...], you should bring it to the attention of the authors [...] However, if the work appears only in a preprint, especially one that is recent and/or not widely cited, you should [...] not penalize them for the missing citation. [...] it is not reasonable to expect a time-consuming empirical comparison with work that has appeared less than 3 months before the submission deadline. In this situation, [...] it should be disregarded when evaluating the novelty of the submission.

Comment chercher des ressources originales ?

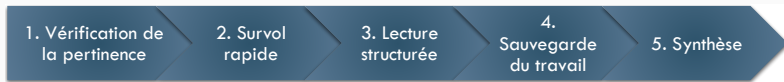
- Pas toujours les mêmes auteurs et labos
 - Couvrir des méthodes et courants de pensée différents
- Rester [curieux.se](#)
 - Des idées inspirantes peuvent venir de là où on ne s'attend pas
 - Aller en conférence, aux groupes de lecture, séminaires
- Petit à petit, constituer une base de connaissances générales

Introduction

Chercher des ressources

Analyser les ressources trouvées

Analyser les ressources trouvées



La structure d'un article

Résumé

- Condensé de l'article, se suffisant à lui-même
- Présente les informations principales

1. Introduction

- Contexte et problématique
- Question de recherche

2. Travaux précédents

- Etat de l'art

3. Méthodologie

- Description de la méthode

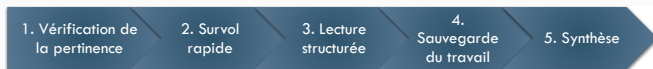
4. Expérimentation & résultats

- Description des expériences réalisées
- Présentation des résultats
- Discussions des résultats et comparaison avec méthodes précédentes

5. Conclusion

- Récapitulatif des travaux présentés et des résultats principaux

1. Vérification de la pertinence de l'article



- Lecture des sections
 - Résumé
 - Introduction
 - Conclusion
- Survol des autres sections
- Le thème est-il pertinent pour mon étude ?
- Les objectifs ont-ils été atteints ?

2. Survol rapide de l'article

1. Vérification de la pertinence

2. Survol rapide

3. Lecture structurée

4. Sauvegarde du travail

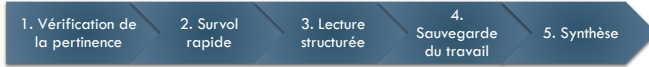
5. Synthèse

- Lecture rapide de l'article **sans s'attarder** sur les points difficiles

Se concentrer sur :

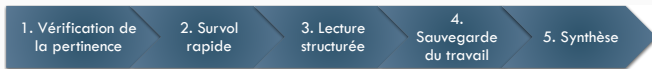
- Résumé, introduction
 - Bien comprendre la problématique
- Figures (images), tableaux de résultats (chiffres) et conclusion
 - Première idée des résultats atteints
- But recherché :
 - Compréhension globale de l'article
 - Peut on le lire en détail tout de suite, ou préparation nécessaire ?

3. Lecture structurée de l'article



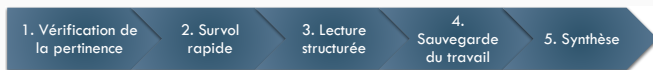
- Lecture détaillée, complète le survol global
- Recherche de ressources annexes si besoin :
 - Articles cités
 - Méthodes citées
 - Travaux précédents des auteur.ice.s
 - ...
- Identifier :
 - But / problématique et **question(s) de recherche**
 - Méthodes utilisées
 - Résultats comparatifs : points forts et faibles
 - **Prise de notes**

4. Sauvegarde du travail



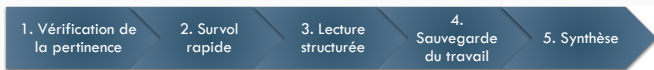
- Liste de ressources
 - Outils de gestion de bibliographie : Zotero, Mendeley, JabRef...
- Notes de lectures
 - Pour chaque article / ressource
 - Format au choix : fichier texte, tableur, papier/crayon...
 - ...
- Rapports synthétiques

4. Sauvegarde du travail



- Citations avec BiBTeX
 - Mise en forme et tri automatiques
 - Disponibles sur la plupart des plate-formes
- Outils de “nettoyage” : `bibclean`
- Conventions pour les identifiants
- Autocomplete sur overleaf et autres éditeurs LaTeX

5. Synthèse

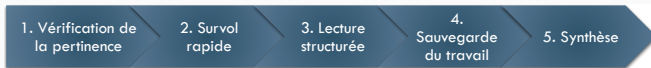


But de la synthèse :

- ~~Empiler articles, les lister indépendamment les uns des autres~~
- Identifier différentes approches pour étudier la question
- Grouper les articles en fonction de caractéristiques communes
 - Théories
 - Méthodologies utilisées
 - Modèles ou algorithmes utilisés
 - Données, datasets
 - Courants de pensée
 - Conclusions obtenues
 - ...

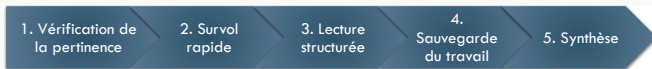
A vous d'identifier la/les caractéristiques pertinentes !

5. Synthèse



- Quoi dire dans une synthèse ?
- Les différentes **catégories** identifiées
- Les articles qui rentrent dans ces catégories
 - Pourquoi ces articles appartiennent (ou pas) aux catégories
 - Quelles **variantes** il apportent
- Analyse critique
 - Éléments de comparaison des différentes catégories
 - Montrer les limites des travaux antérieurs : aspects manquants ou insatisfaisants pour notre problématique

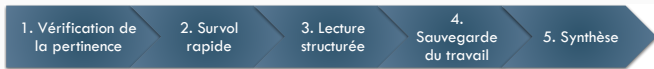
5. Synthèse



Méthode :

1. Lire les articles en prenant des notes
2. Choisir la ou les **catégorisation(s)** à utiliser
 - Possibilité de les organiser par progression logique (ex. deep learning, analyse syntaxique en dépendance, puis multilinguisme, ...)
3. Identifier les **sous-catégories / variantes** principales (ex. réseaux récurrents, réseaux bidirectionnels, réseaux adversaires, ...)
4. Présenter les articles de chaque catégorie et sous-catégorie
 - Avec citation '[x]' ou '(Untel 2016)'

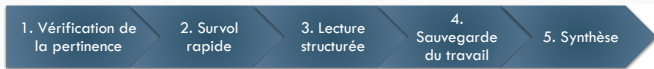
5. Synthèse



Pour écrire une bonne synthèse :

- Pas un catalogue d'articles sans liens entre eux
 - En 1-2 phrases : résumé de ce qui est important dans l'article
 - Focus : aspects pertinents pour mon travail (similaire/différent)
 - S'aider d'outils d'aide à la rédaction : chatGPT, ...

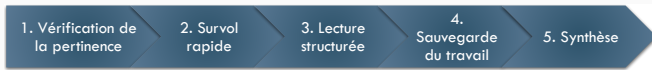
5. Synthèse



Pour écrire une bonne synthèse :

- Pas un catalogue d'articles sans liens entre eux
 - En 1-2 phrases : résumé de ce qui est important dans l'article
 - Focus : aspects pertinents pour mon travail (similaire/différent)
 - S'aider d'outils d'aide à la rédaction : chatGPT, ...
- Rester centré sur la problématique
- Donner l'**essentiel** de l'information
 - Pas trop de détails sur chaque article
 - Ce n'est pas un cours et nous parlons à des expert.e.s

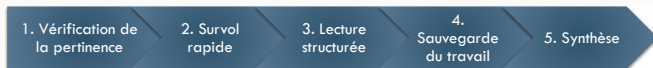
5. Synthèse



Pour écrire une bonne synthèse :

- Savoir à qui on s'adresse
 - En général : **collègues** du même domaine de recherche

5. Synthèse



Pour écrire une bonne synthèse :

- Savoir à qui on s'adresse
 - En général : **collègues** du même domaine de recherche
- Toujours indiquer ses **sources**
- Citer des sources "respectées"
 - Publications scientifiques, rapports officiels...
 - Attention aux articles trop récents, non relus (arXiv)

5. Synthèse : exemple à éviter

CITATION1 a montré que XXX. CITATION2 a utilisé Y1 et ont montré que cette technique est la plus efficace.

Y2 a été utilisé par CITATION3 qui ont mis en évidence AVANTAGE. CITATION4 ont préféré utiliser Y1, tout comme CITATION2 et CITATION5.

5. Synthèse : exemple

Ce travail est à l'intersection de trois tendances dans la littérature sur l'YY. La première est le Y1, qui consiste à XXX. L'avantage de cette méthode est que XXX. La seconde est Y2, qui a pour but de XXX. Cette deuxième méthode à l'avantage de XXX. La troisième et dernière tendance est l'utilisation de Y3, qui offrent de nombreux renseignements sur XXX.

Le Y1 est une solution efficace lorsque XXX. CITATION décrivent deux types de Y1. Le Y1 repose sur les AVANTAGES. Par exemple, CITATION proposent XXX. Ces méthodes donnent parfois des résultats inattendus : XXX. L'hypothèse des auteurs serait que XXX.

Comme nous l'avons mentionné ci-dessus, les Y2 utilisent XXX. L'approche proposée par CITATION consiste à adapter un XXX. Cette méthode peut être utilisée pour XXX. CITATION montrent que les Y2 peuvent être XXX, donnant de meilleurs résultats que des YY.

Une autre source d'information pouvant être utilisée est Y3, tel que ceux présents dans le CITATION ou CITATION qui donnent des informations sur XXX. CITATION et CITATION utilisent l'ensemble des XXX. Ces AVANTAGES, que nous définirons et utiliserons également par la suite, sont les suivants : XXX. CITATION utilisent plutôt YY car DESAVANTAGE de Y3.

5. Synthèse : exemple à éviter

CITATION provides XXX as context in their dataset. CITATION, CITATION and CITATION also provide XXX as context. Subjectivity is addressed by CITATION using METHOD. However, CITATION reach a performance of XXX and CITATION reach a performance of. Sometimes, it is possible to also use METHOD (CITATION). The most common solution is XXX (CITATION, CITATION).

5. Synthèse : exemple

The amount of context available for XXX varies : most dataset contain XXX (CITATION,CITATION,CITATION), but some of them provide larger contexts such as XXX (CITATION), or even XXX (CITATION). The resulting XXX are often subjective, because EXPLANATION. One way to attenuate this subjectivity is to XXX, assuming that HYPOTHESIS (CITATION). In addition, it is possible to guide annotators by METHOD (CITATION), or to provide INFORMATION (CITATION). Some authors, however, argue that they prefer not biasing annotators, so XXX (CITATION).

Question : *Choix méthodologiques faits dans les expériences ?*

Portée

- Articles >2016 utilisant datasets X ou Y, disponibles sur Z

Grille d'analyse

- 2. Données
 - 2.1 Source (news, wikipédia, réseaux sociaux...)
 - 2.2 Langues
- 3 - Traitement des données
 - 3.1 Pre-traitements
 - 3.2 Post-traitements
- 4 - Évaluation
 - 4.1 Métriques (exact match, fuzzy match...)
 - 4.2 Tests de significativité

Source: Inspiré de <https://aclanthology.org/2023.mwe-1.15/>

Revue de la littérature : exemple

	2 Languages	3 Split of the corpora	3.4 Category of	4.1 Preprocess
PARSEME 1.0				
The PARSEME Shared Task on Autom	18: BG, CS, DE, EL,	train/test, no dev		N/A
Parsing and MWE Detection: Fips at th	8: FR, EN, DE, IT, E	Not mentioned	VID, LVC, VPC,	Transformation
The ATILF-LLF System for Parseme Sh	18: BG, CS, DE, EL,	PARSEME data	PARSEME cate	Not mentioned
Detection of Verbal Multi-Word Express	15: CS, DE, EL, ES,	PARSEME data	VPC, LVC, VID,	Not mentioned
USzeged: Identifying Verbal Multiword	9: DE, EL, ES, FR, F	PARSEME 1.0 (no dev	PARSEME 1.0 c	Remove long se
A data-driven approach to verbal multy	12: RO, FR, CS, DE	PARSEME 1.0 - cross	PARSEME 1.0 c	Not mentioned
Neural Networks for Multi-Word Express	15: BG, CS, DE, EL,	80% train, 10% dev, 10	PARSEME 1.0	Not mentioned
PARSEME 1.1				
Edition 1.1 of the PARSEME Shared Ta	19: BG, DE, EL, EN,	3 languages had no de	LVC, VID, IRV, V	N/A
CRF-Seq and CRF-DepTree at PARSE	19: BG, DE, EL, EN,	PARSEME 1.1 data	PARSEME 1.1	Converting to X
Deep-BGT at PARSEME Shared Task	10: BG, DE, ES, FR,	PARSEME 1.1 data	All PARSEME 1.	Merging labels, l

- Faire attention aux potentiels biais de l'étude
- Discuter avec les collègues si quelque chose semble "bizarre"
 - Si on n'a pas compris, c'est peut-être juste pas clair
 - Un langage peu clair peut cacher une méthodologie pas nette
 - Ne pas avoir peur d'avoir l'air bête – ça permet d'avancer
 - Écrire aux auteur.ice.s, demander leur code, données si possible
- Publish or perish : des relecteur.ice.s pressé.e.s peuvent laisser passer des articles problématiques

Exemple

Ganley, Mingle, Ryan, Ryan, Vasilyeva, Perry (2013). Developmental Psychology <https://psycnet.apa.org/record/2013-02693-001>

- [...] no evidence that the mathematics performance of school-age girls was impacted by stereotype threat.
- Condition expérimentale “stéréotype” : *This is very important, as boys have done much better than girls on this test in the past.*
- Condition expérimentale “non stéréotype” : on ne dit rien
→ Or, le stéréotype est aussi activé en l'absence de la consigne !

Source : formation “inégalités de genre” - Isabelle Régner et Magali Putero

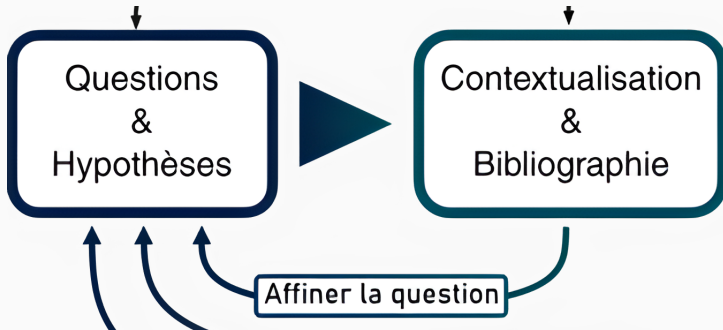
Une proposition de structuration du travail :

- **Titre, auteur**
- **Résumé**
- **Contributions**
- **Similarités** avec vos travaux
- **Différences** avec vos travaux
- **Remarques**

“Notre” framework : bibliographie

Paper	Résumé	Contributions	Similarités	Différence	Remarques
Wang and Eisner (2018)	Cet article parle de l'utilisation de statistiques de surface pour faire du parsing sur une langue inconnue ...	<ul style="list-style-type: none"> - Les features apprises sur des corpus annotés en POS aident au parsing - L'utilisation de langues "synthétiques" au moment de l'entraînement augmente les résultats - ... 	<ul style="list-style-type: none"> - Delexicalisé - Zero-shot - ... 	<ul style="list-style-type: none"> - Non supervisé - Ils n'utilisent aucune donnée parallèle - ... 	<ul style="list-style-type: none"> - Leur système dépend des POS tags gold - Critiques intéressantes du WALS : <ul style="list-style-type: none"> - <i>The unknown language might not be in WALS</i> - ...
de Lhoeneux et al. (2018)	Le but de cet article est de tester différentes stratégies de partage des paramètres d'un parser en dépendance. Ils testent ...	<ul style="list-style-type: none"> - Certains paramètres sont utiles au partage de connaissances, d'autres non - Le MLP qui sert de classifieur - ... 	<ul style="list-style-type: none"> - Multilingue - Zero-shot - ... 	<ul style="list-style-type: none"> - Lexique - Semi-supervisé - ... 	<ul style="list-style-type: none"> - Ils testent 27 configurations différentes - Leur tests se font en entraînant 5 modèles bilingues - ...
...

Rappel : processus itératif



- Première étape de tout projet de recherche
- Processus itératif - définition de la question de recherche
- Objectifs
 - Construire un **argumentaire scientifique**
 - **Justifier** la question de recherche
 - Pertinente - combler une lacune
 - Faisable - s'inspirer de ce qui existe
 - Intéressante - impact potentiel dans le domaine

- Synthèse de l'état de l'art dans le domaine
- Nécessite une lecture structurée d'un grand nombre d'articles
 - Capacité de synthèse
 - Comparaison, mise en perspective
 - Analyse et structuration des connaissances
 - Identification des défis et problèmes ouverts du domaine
- Valorisée / publiée sous la forme d'un survey ou meta-analyse

- Cours d'Adeline Paiement
- Cours de méthodologie de Damien Driot
- Google images
- Google Bard
- ChatGPT