



Building a literature review

Karën Fort

karen.fort@univ-lorraine.fr / <https://members.loria.fr/KFort>

Sources of inspiration

- ▶ Introduction to the Research Process: Tools and Methodology, Stéphane Zuckerman and Jordane Lorandel, Université de Cergy-Pontoise
- ▶ Wrestling an elephant into a cupboard: how to write a PhD literature review in nine easy steps

What do you know about it?

from your M1 project



Motivations

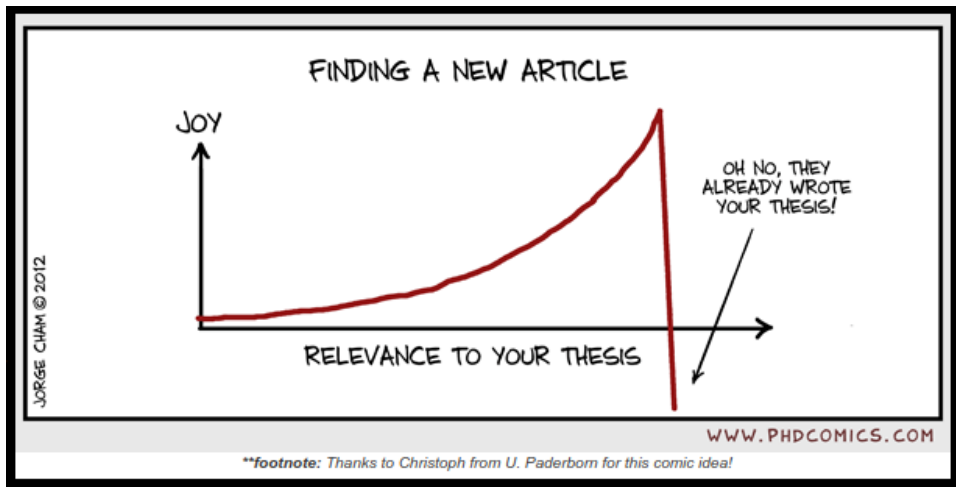
What are we talking about

Where to look

How to read a paper

Why doing it?

to avoid wasting (your) time



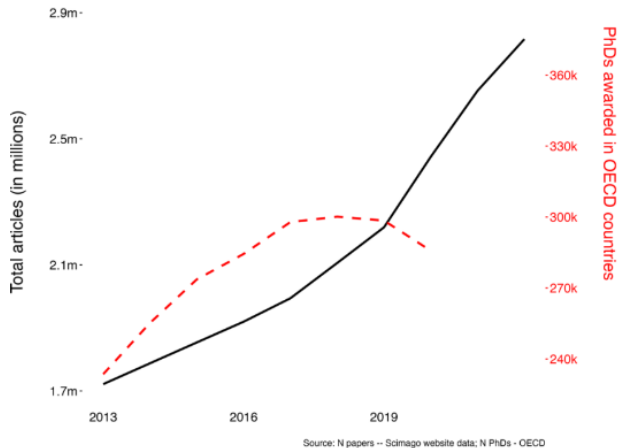
all images © jorge cham



<https://phdcomics.com/comics.php?f=1506>

Why doing it?

to avoid wasting (others') time



[Hanson et al., 2024]

Why doing it?

to compare yourself

Influence of Pre-annotation on POS-tagged Corpus Development

Karën Fort

INIST CNRS / LIPN

Nancy / Paris, France.

`karen.fort@inist.fr`

Benoît Sagot

INRIA Paris-Rocquencourt / Paris 7

Paris, France.

`benoit.sagot@inria.fr`

[Fort and Sagot, 2010]

Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation

Ines Rehbein and **Josef Ruppenhofer** and **Caroline Sporleder**

Computational Linguistics

Saarland University

`{rehbein, josefr, csporled}@coli.uni-sb.de`

[Rehbein et al., 2009]

Why doing it?

to renew old ideas

with which they are actually found. More than that: if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

<https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520>

[Harris, 1954]

Why doing it?

Not only because you do not want to:

- ▶ re-invent the wheel
- ▶ participate to the exponential growth of published papers

But also to:

- ▶ better identify the strengths of your work
- ▶ be inspired

and yes, NLP did not start existing in 2013!

Motivations

What are we talking about

Where to look

How to read a paper

What a (PhD) literature review isn't

"A PhD literature review isn't just a summary of existing literature"

Wrestling an elephant into a cupboard: how to write a PhD literature review in nine easy steps

What a (PhD) literature review is

*"A PhD literature review is a **critical assessment** of the literature in your field and related to your specific research topic. When discussing each relevant piece of literature, the review must **highlight where the gaps are and what the strengths and weaknesses are** of particular studies, papers, books, etc. Also, different pieces of literature are **compared and contrasted** with one another so that themes and relationships are highlighted."*

Wrestling an elephant into a cupboard: how to write a PhD literature review in nine easy steps

Motivations

What are we talking about

Where to look

How to read a paper

Where to start (easy)

- ▶ seed: your advisor
- ▶ places to search:
 - ▶ the ACL anthology: <https://aclanthology.org/> – more later
 - ▶ esp. specialized workshops
 - ▶ State of the art / literature reviews in PhD theses
 - ▶ Semantic Scholar: <https://www.semanticscholar.org/> – more later
 - ▶ your favorite search engine
 - ▶ National projects, EU projects

Beware:

- ▶ use arXiv only near the end (lot of noise)
- ▶ social networks: can help, but biased view (hype)

Paper vs paper

Exercise

what is the difference between an arXiv paper and a paper on the ACL anthology?

About the ACL anthology: taking shortcuts



The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research

Mohamed Abdalla | Jan Philip Wahle | Terry Ruas | Aurélie Névéol | Fanny Duclé | Saif Mohammad | Karen Fort

About Semantic Scholar: snowballing

SEMANTIC SCHOLAR

amazon mechanical turk gold mine

Search

Sign In

Create Free Account

DOI: 10.1162/COLL_a_00057 • Corpus ID: 1051130

Last Words: Amazon Mechanical Turk: Gold Mine or Coal Mine?

Karên Fort, G. Adda, K. Cohen • Published in *International Conference on...* • Computer Science

Share

372 Citations

- Highly Influential Citations 13
- Background Citations 174
- Methods Citations 13
- Results Citations 9

View All

View on ACL

PDF | direct.mit.edu

Save to Library

Create Alert

Cite

Going further (less easy)

- ▶ find the **right terminology** (takes time)
- ▶ find the **key actors**
- ▶ reach towards other disciplines/fields

Getting a paper

What if you found a paper, but it's under a paywall?

- ▶ try to find it elsewhere: HAL, arXiv, authors' page
- ▶ write to the authors
- ▶ you **cannot** use SciHub. It's bad bad bad.

Motivations

What are we talking about

Where to look

How to read a paper

Beware: it takes time to read a (good) paper...



all images © jorge cham



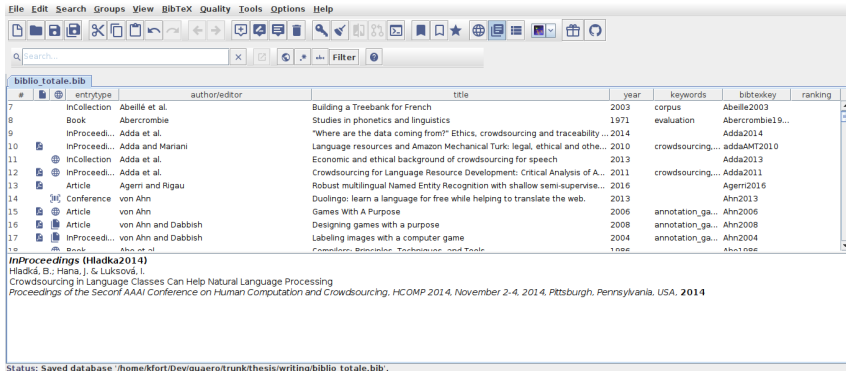
<https://phdcomics.com/comics/archive.php?comid=984>

... and reading is not enough

You absolutely need to:

- ▶ **annotate/take notes** in the papers or in a tool
- ▶ **store** the bibtex entries (all of them) (and the papers, if needed)
- ▶ **tag/index** the papers you read

Tools can help you



The screenshot shows the JabRef application window. The title bar reads "biblio_totale.bib". The menu bar includes "File", "Edit", "Search", "Groups", "View", "BibTeX", "Quality", "Tools", "Options", and "Help". Below the menu is a toolbar with various icons for file operations, search, and editing. A search bar is located below the toolbar. The main area displays a table of bibliographic entries with columns for "#", "entrytype", "author/editor", "title", "year", "keywords", "bibtexkey", and "ranking".

#	entrytype	author/editor	title	year	keywords	bibtexkey	ranking
7	InCollection	Abellé et al.	Building a Treebank for French	2003	corpus	Abelle2003	
8	Book	Abercrombie	Studies in phonetics and linguistics	1971	evaluation	Abercrombie19...	
9	InProceedi...	Adda et al.	"Where are the data coming from?" Ethics, crowdsourcing and traceability ...	2014		Adda2014	
10	InProceedi...	Adda and Mariani	Language resources and Amazon Mechanical Turk: legal, ethical and othe...	2010	crowdsourcing...	addaAMT2010	
11	InCollection	Adda et al.	Economic and ethical background of crowdsourcing for speech	2013		Adda2013	
12	InProceedi...	Adda et al.	Crowdsourcing for Language Resource Development: Critical Analysis of A...	2011	crowdsourcing...	Adda2011	
13	Article	Agerri and Rigau	Robust multilingual Named Entity Recognition with shallow semi-supervise...	2016		Agerri2016	
14	Conference	von Ahn	Duolingo: learn a language for free while helping to translate the web.	2013		Ahn2013	
15	Article	von Ahn	Games With A Purpose	2006	annotation_ga...	Ahn2006	
16	Article	von Ahn and Dabbish	Designing games with a purpose	2008	annotation_ga...	Ahn2008	
17	InProceedi...	von Ahn and Dabbish	Labeling images with a computer game	2004	annotation_ga...	Ahn2004	
18	Book	Ahn et al.	Compend: Principles, Techniques, and Tools	1988		Ahn1988	

InProceedings (Hladka2014)
Hladká, B.; Hana, J. & Luksová, I.
Crowdsourcing in Language Classes Can Help Natural Language Processing
Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA, 2014

Status: Saved database '/home/kfort/Dev/quaero/trunk/thesis/writing/biblio_totale.bib'.

JabRef

Practice

JabRef vs Zotero?

What is a proper bibtex entry?

for a conference paper

```
@InProceedings{Bird2020,  
  author      = {Bird, Steven},  
  title       = {Decolonising Speech and Language Technology},  
  booktitle   = {Proceedings of the 28th International Conference on Computational  
                Linguistics},  
  year        = {2020},  
  pages       = {3504--3519},  
  address     = {Barcelona, Spain (Online)},  
  month       = {December},  
  publisher   = {International Committee on Computational Linguistics},  
  url         = {https://www.aclweb.org/anthology/2020.coling-main.313},  
}
```

What is a proper bibtex entry?

for a journal paper

```
@Article{Artstein2008,  
  Title      = {Inter-Coder Agreement for Computational Linguistics},  
  Author     = {Artstein, Ron and Poesio, Massimo},  
  Journal    = {Computational Linguistics},  
  Year       = {2008},  
  Number     = {4},  
  Pages      = {555--596},  
  Volume     = {34},  
  Address    = {Cambridge, MA, USA},  
  Doi        = {http://dx.doi.org/10.1162/coli.07-034-R2},  
  File       = {:/home/kfort/Dev/quaero/trunk/thesis/writing/Articles_biblio/  
               IAAgreement.pdf:PDF},  
  ISSN       = {0891-2017},  
  Publisher  = {MIT Press},  
  Url        = {http://www.mitpressjournals.org/doi/abs/10.1162/coli.07-034-R2}  
}
```

What is a proper bibtex entry?

for an arXiv paper

```
@Misc{Hanson2024,  
  author      = {Mark A. Hanson and Pablo Gómez Barreiro and Paolo Crosetto  
                and Dan Brockington},  
  title       = {The strain on scientific publishing},  
  year        = {2024},  
  archiveprefix = {arXiv},  
  eprint      = {2309.15884},  
  primaryclass = {cs.DL},  
  url         = {https://arxiv.org/abs/2309.15884},  
}
```





Be (pro-)active in your team

- ▶ participate to reading groups
- ▶ organize one

Tricky question

Practice

Find papers about ethics in NLP that were published before 2016

-  Fort, K. and Sagot, B. (2010).
Influence of pre-annotation on POS-tagged corpus development.
In Proceedings of the Fourth ACL Linguistic Annotation Workshop, pages 56–63,
Uppsala, Suède.
-  Hanson, M. A., Barreiro, P. G., Crosetto, P., and Brockington, D. (2024).
The strain on scientific publishing.
-  Harris, Z. (1954).
Distributional structure.
Word, 10(23):146–162.
-  Rehbein, I., Ruppenhofer, J., and Sporleder, C. (2009).
Assessing the benefits of partial automatic pre-labeling for frame-semantic
annotation.
In Proceedings of the Third Linguistic Annotation Workshop, pages 19–26,
Singapour. Association for Computational Linguistics.