Ethics and Natural Language Processing (NLP): an (almost) tutorial

Karën Fort

karen.fort@loria.fr / https://members.loria.fr/KFort/

ICMI, Oct. 18th, 2021

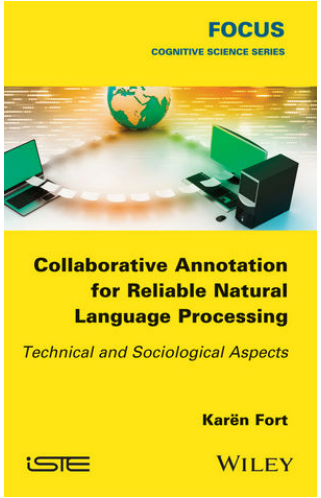
Loria
Laboratoire lorrain de recherche en informatique et ses applications

LETTRES SORBONNE UNIVERSITÉ

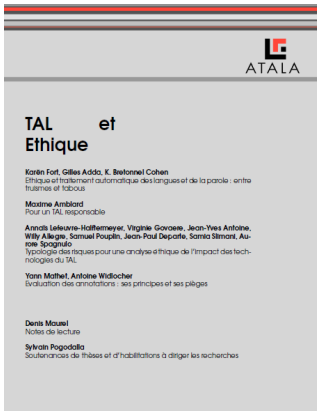
Where I'm talking from

See https://members.loria.fr/KFort/

- Language resources creation for NLP, esp. using crowdsourcing



- Ethics and NLP



We all use NLP, every day

Google

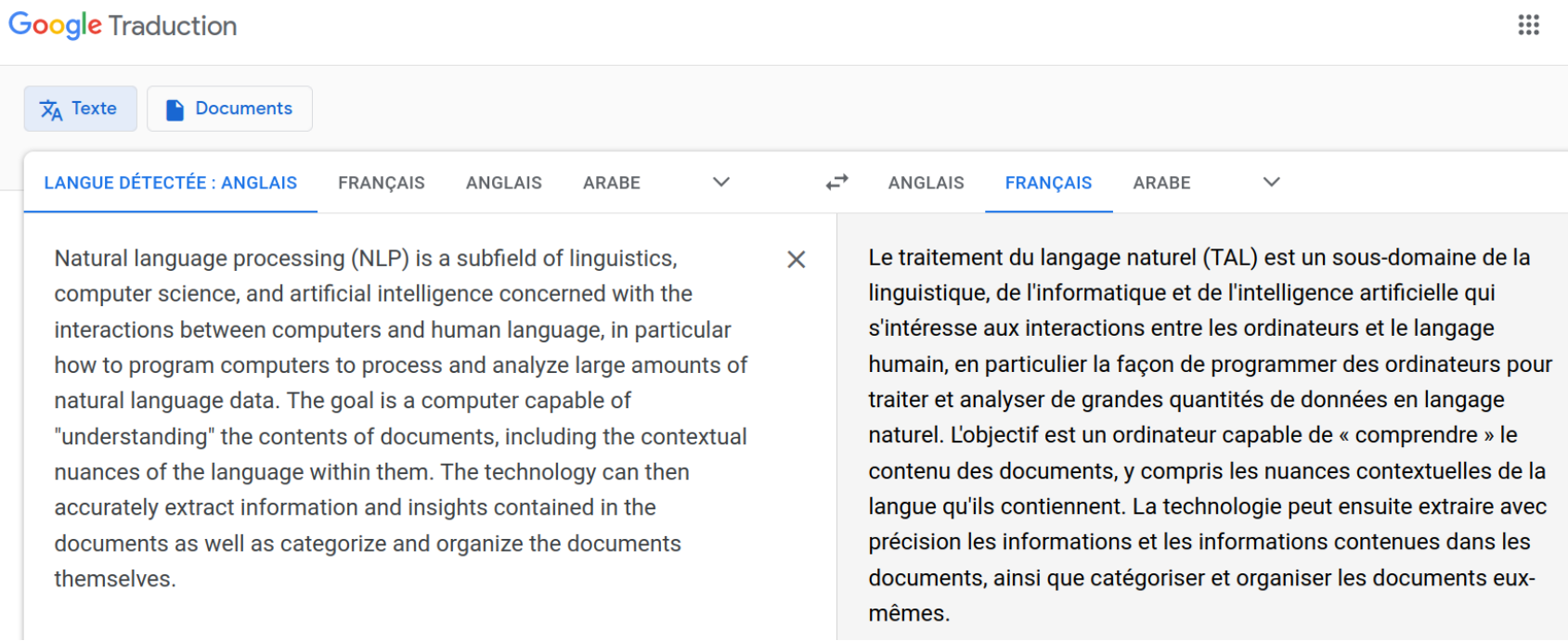
natural language processing is

- natural language processing is **a type of**
- natural language processing is **divided into the two subfields of**
- natural language processing is **ai that**
- natural language processing is **a component of**
- natural language processing is **used in**
- natural language processing is **a field of**
- natural language processing is **a type of quizlet**
- natural language processing is **associated with which of the followin g areas**
- natural language processing is **used in mcq**
- natural language processing is **all that**

https://www.google.com/

What is NLP?

A recursive definition



https://translate.google.com/

(More advanced) applications in our daily lives

. . . not necessarily very efficient

The SNCF (French railway) chatbot on https://twitter.com/home

Two revolutions in less than a decade

- Much more implication from big firms in the field (GAFAM)
- Deep learning (from approx. 2013), including (very) large language models

$\rightarrow$ ethical issues start to show

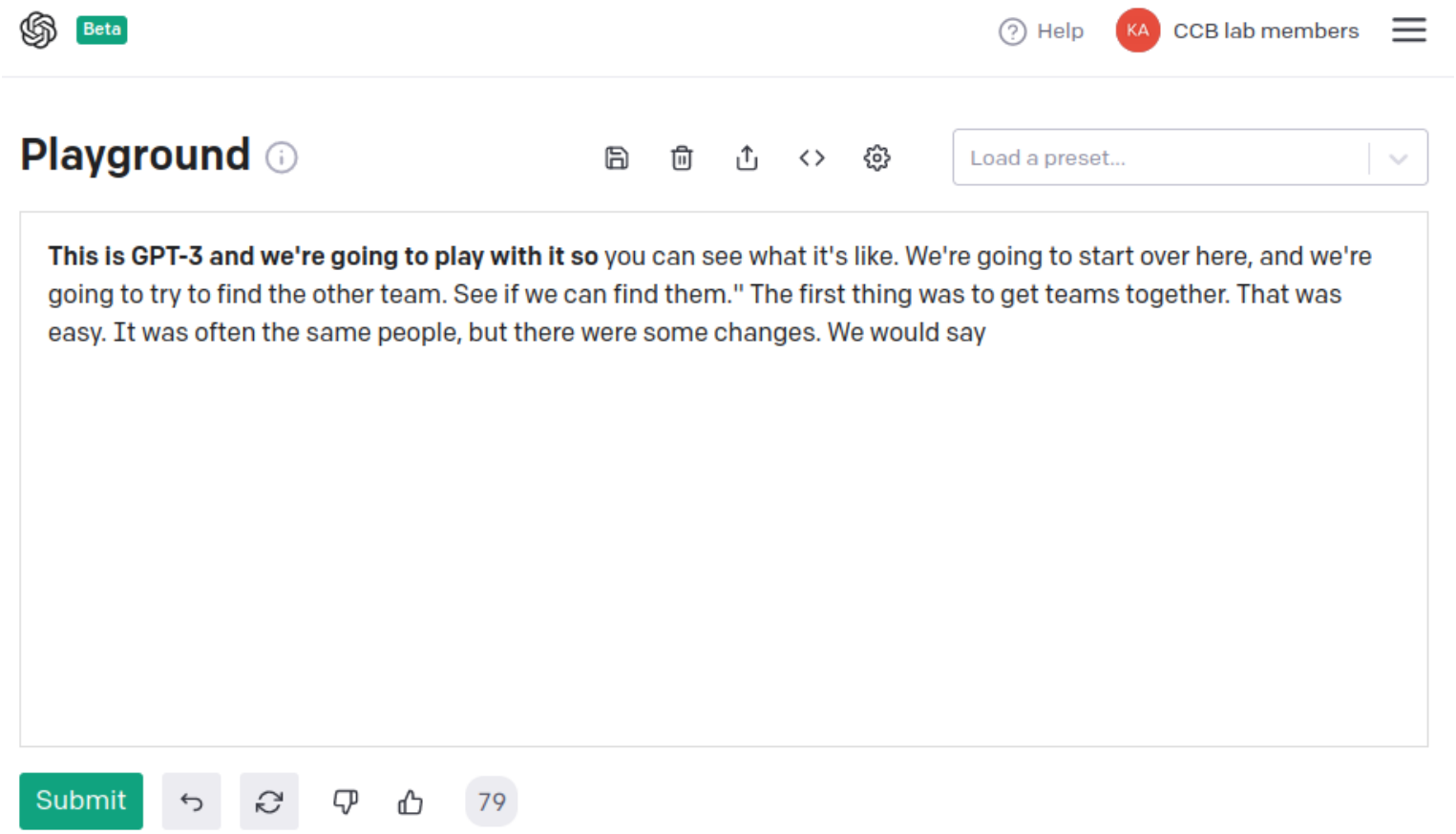
Who is talking, from where?

Practicing ethics in NLP

Helping the thought

Exercise: let's prompt!

Thanks to Chris Callison-Burch!

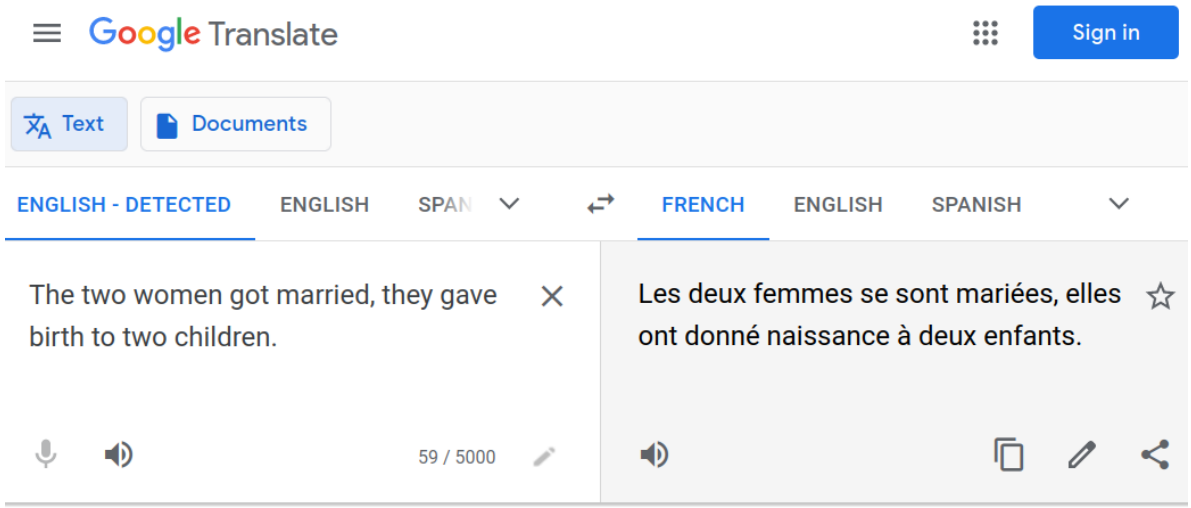


https://beta.openai.com/playground

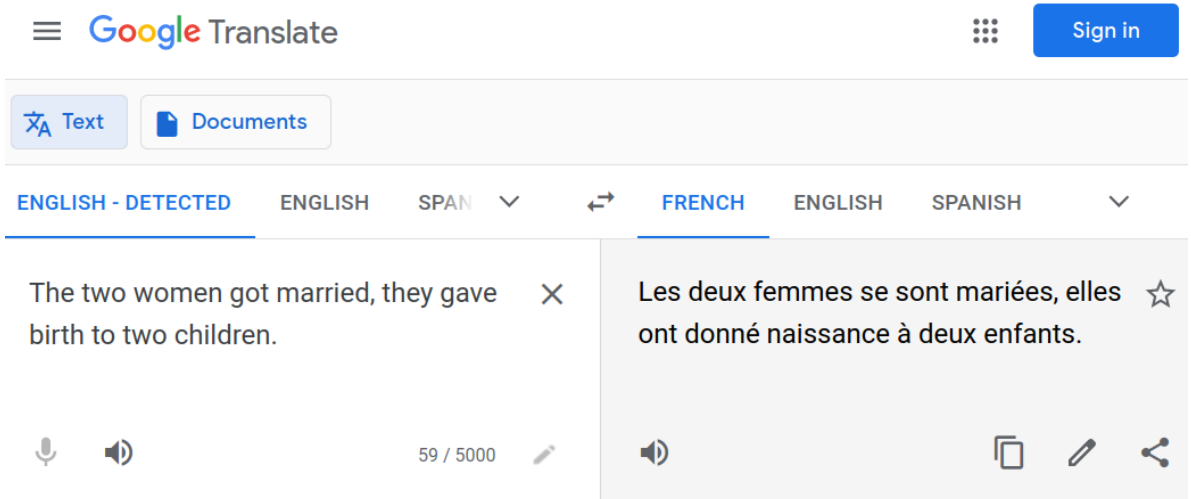
Exercise: which ethical issues did you find?

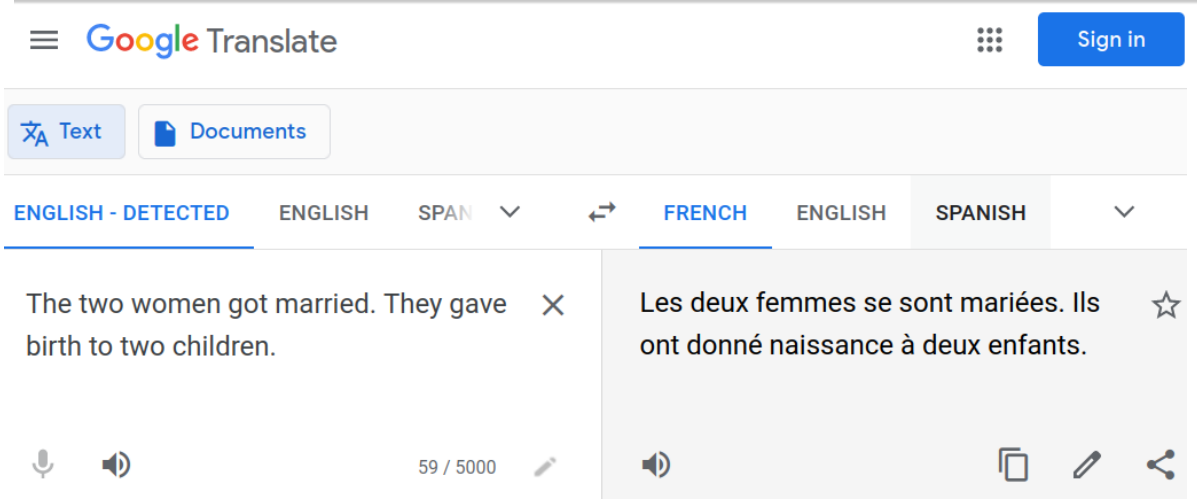
Exercise: where do the stereotypes come from?

Mirror of existing stereotypes?

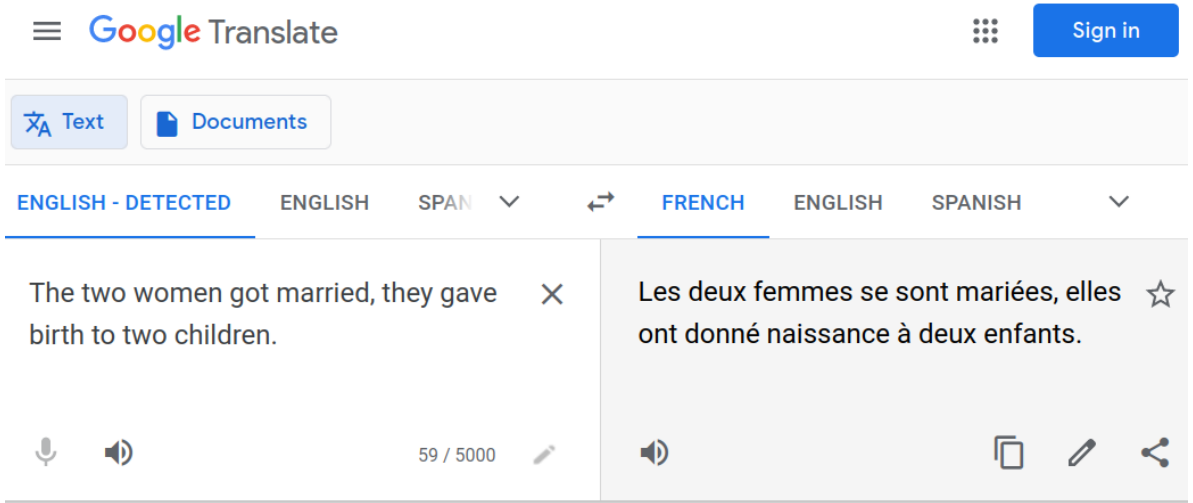


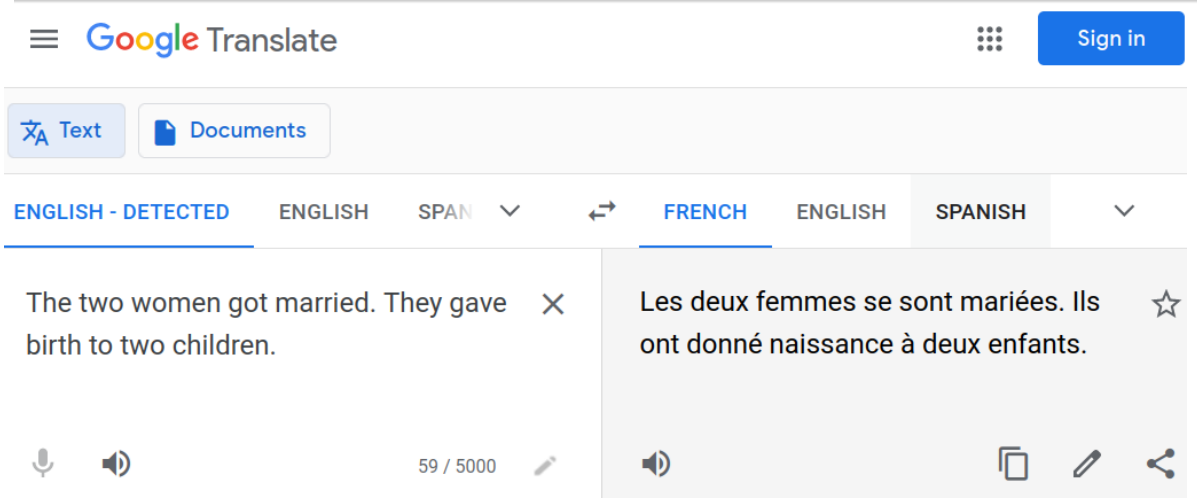
Mirror of existing stereotypes?





Mirror of existing stereotypes?





context taken into account (sentence) + masculine = neutral

Machine learning is not magic

The decisions to:

- define masculine as neutral in French (not the case in Ancient French)
- take the sentence as the context

were MADE by people

ML and biases: mirror or amplifier?

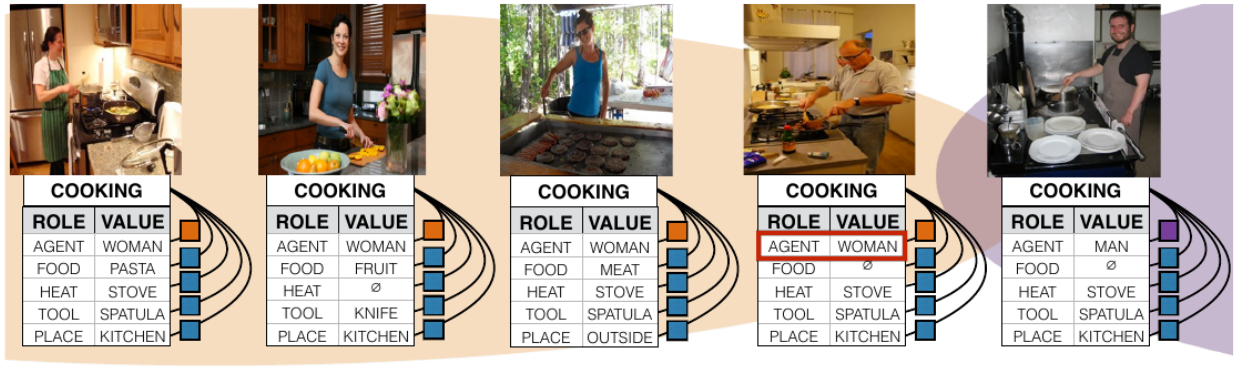


Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, `cooking`, its semantic roles, i.e `agent`, and noun values filling that role, i.e. `woman`. In the imSitu training set, 33% of `cooking` images have `man` in the `agent` role while the rest have `woman`. After training a Conditional Random Field (CRF), bias is amplified: `man` fills 16% of `agent` roles in `cooking` images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, `man` appears in the `agent` role of 20% of `cooking` images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

[Zhao et al., 2017]

(At least some) hype benefits ethics

[Hovy and Spruit, 2016] (mainly) about biases in NLP:

(At least some) hype benefits ethics

[Blodgett et al., 2020] analyzed 146 articles about biases in NLP:

Exercise: any idea about the environmental impact of NLP?

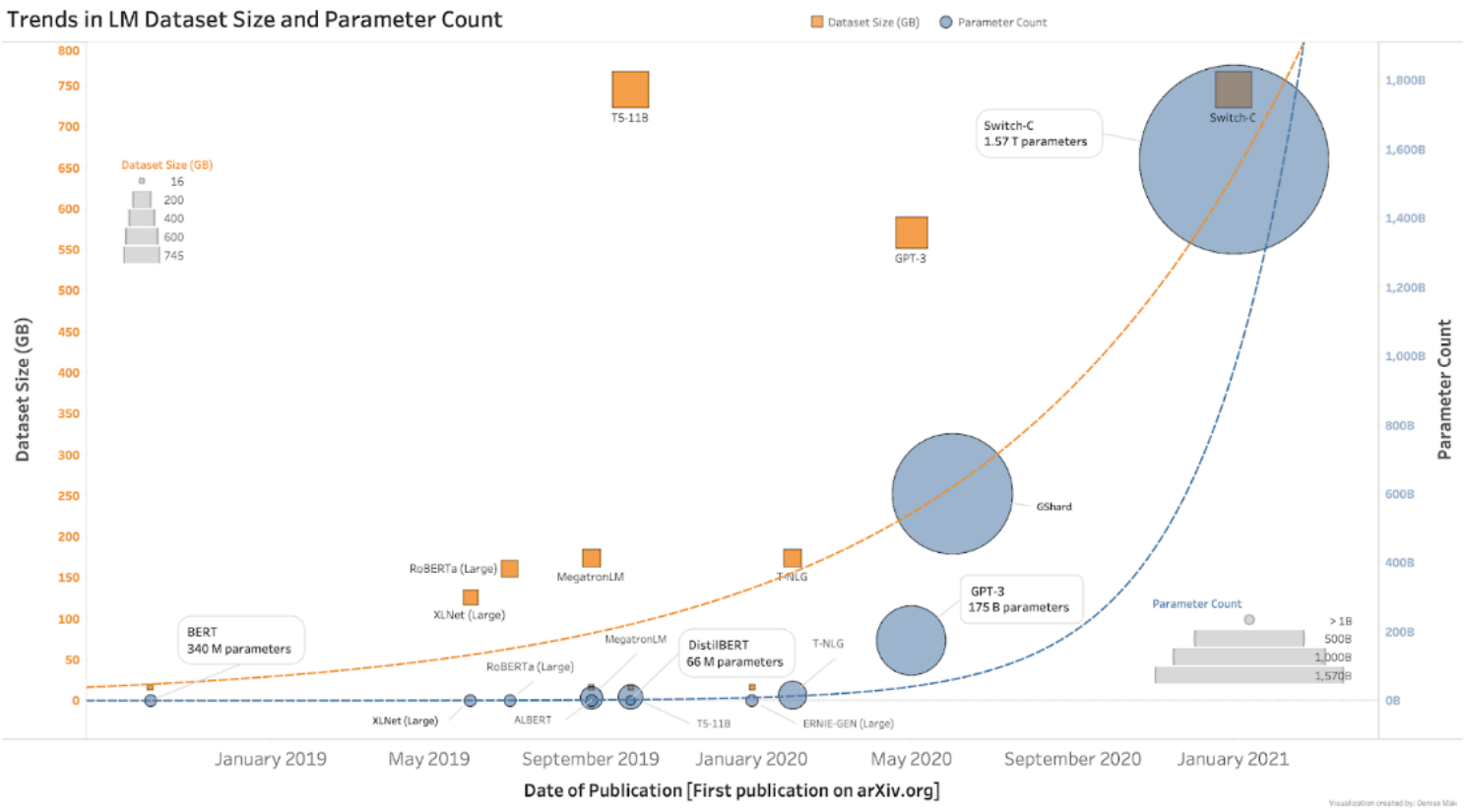
Carbon footprint

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

[Strubell et al., 2019]

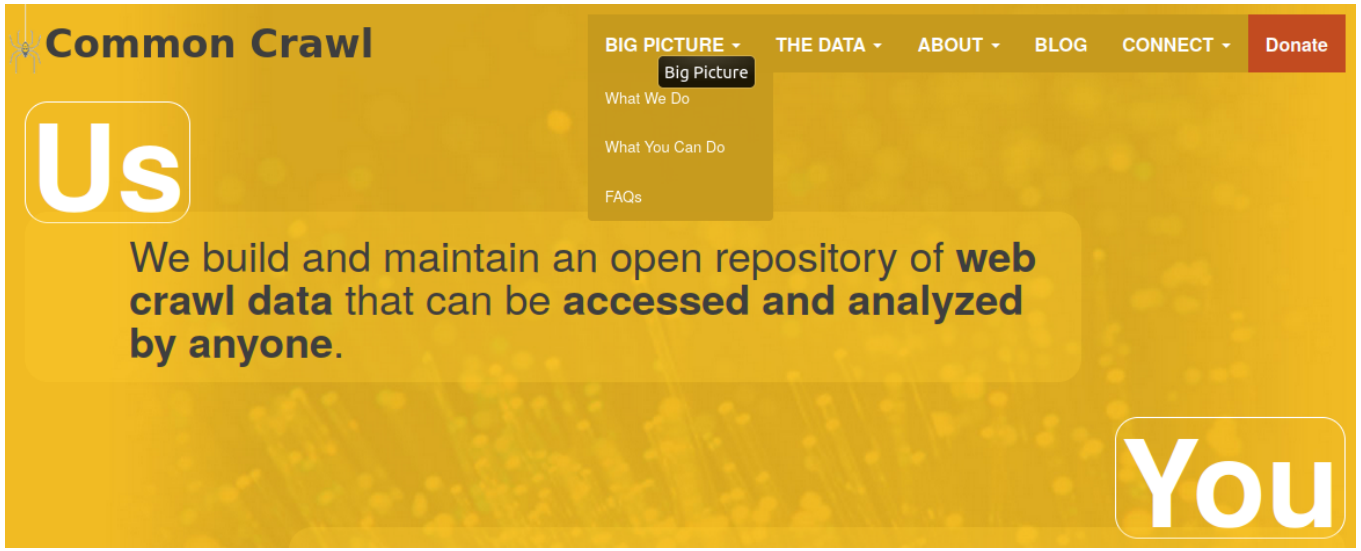
Models trained once and for all?



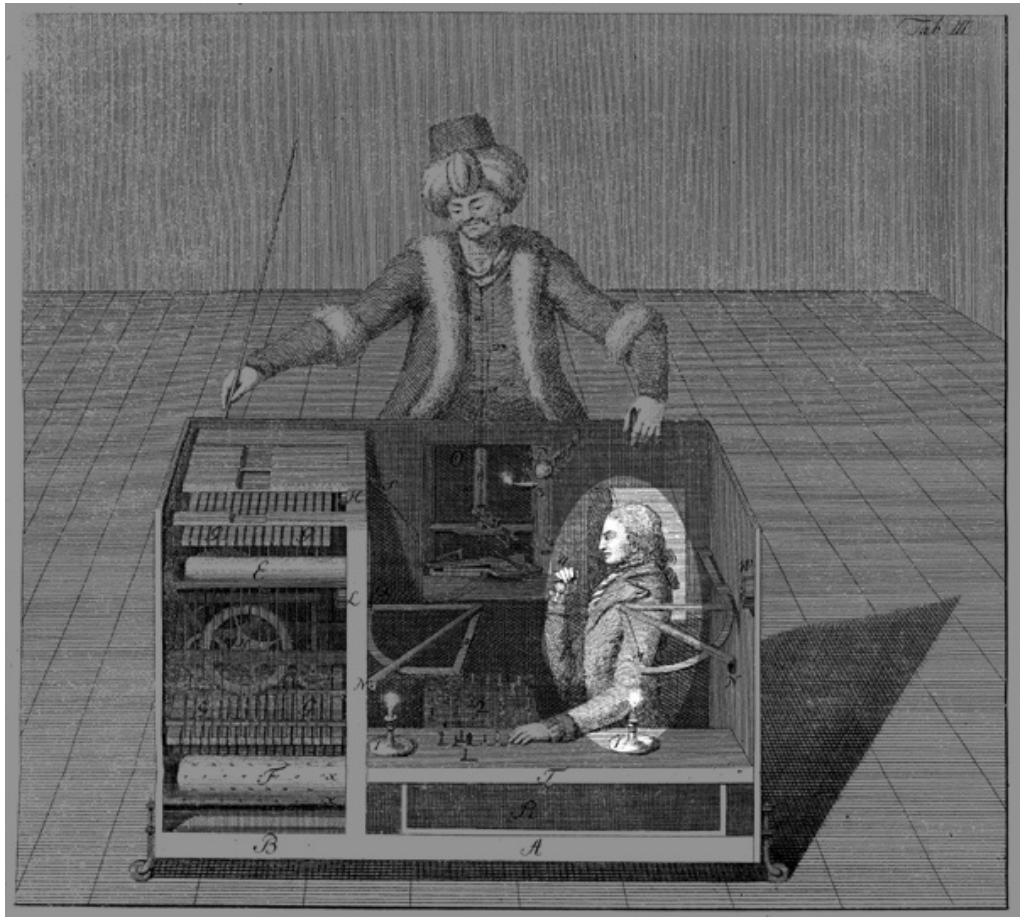
[Bender et al., 2021]

Exercise: where does data come from?

Data and "informed" consent



Artificial artificial intelligence



[Fort et al., 2011]

Data not "for the taking"

Decolonising Speech and Language Technology

Steven Bird
Northern Institute
Charles Darwin University

Abstract

After generations of exploitation, Indigenous people often respond negatively to the idea that their languages are data ready for the taking. By treating Indigenous knowledge as a commodity, speech and language technologists risk disenfranchising local knowledge authorities, reenacting the causes of language endangerment. Scholars in related fields have responded to calls

[Bird, 2020]

"Overselling" research results

Accueil > Espace presse

Invitation à la journée « Intelligence artificielle : l'ordinateur passe la barrière de la langue »

04 janvier 2021 NUMÉRIQUE

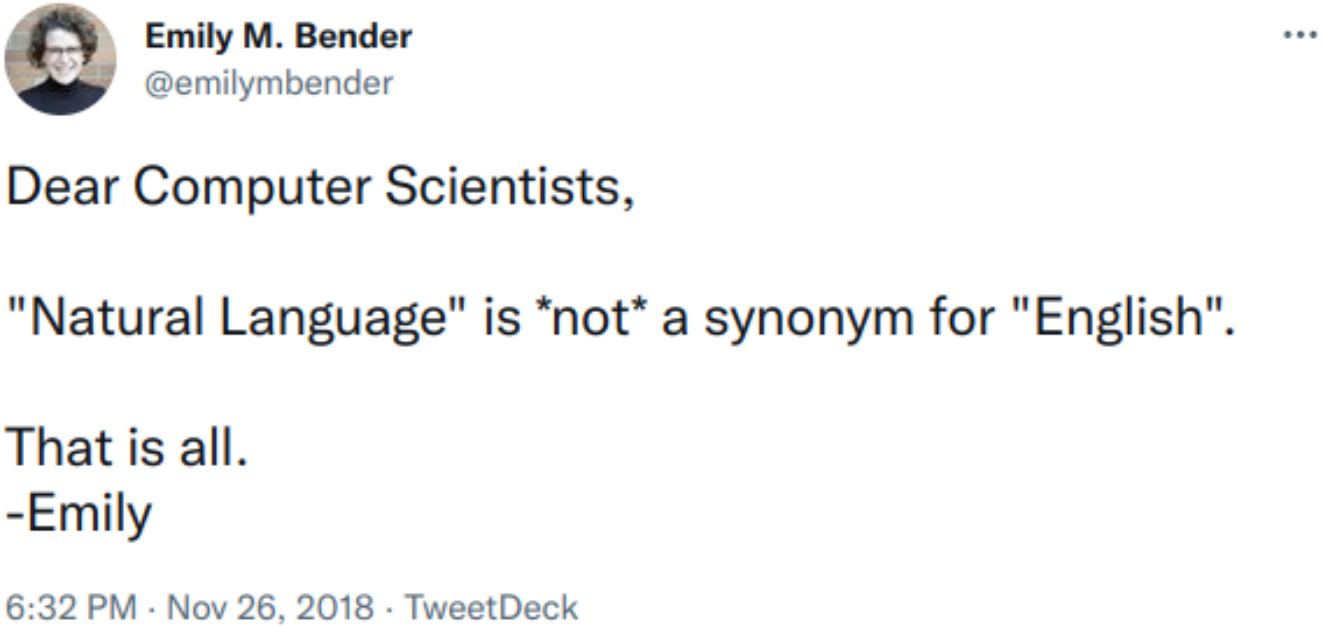
vs [Bender and Koller, 2020]

Climbing towards NLU:
On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

By the way. . .



https://twitter.com/emilymbender/status/1067108757488848896

Evaluation issues

Pratiques d'évaluation en ASR et biais de performance

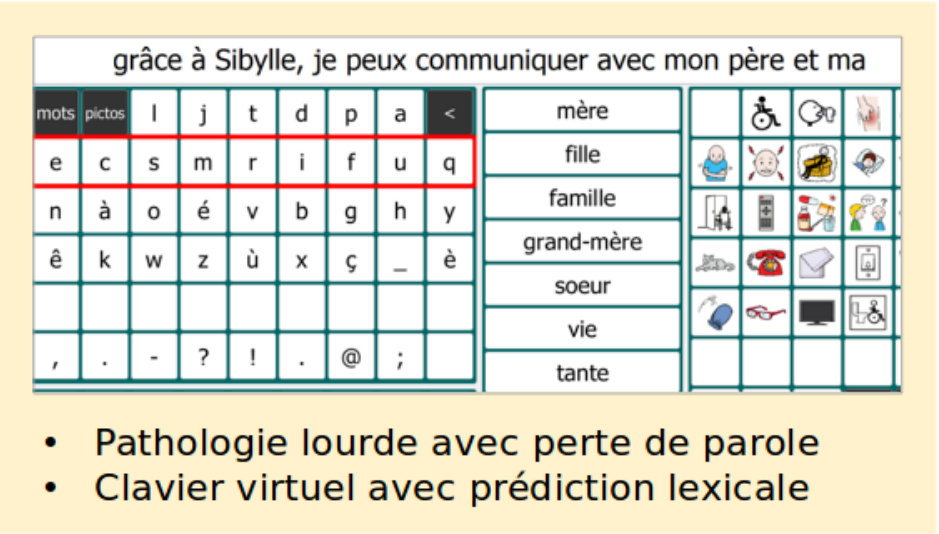
Mahault Garnerin[1, 2] Solange Rossato[2] Laurent Besacier[2]
(1) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
(2) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

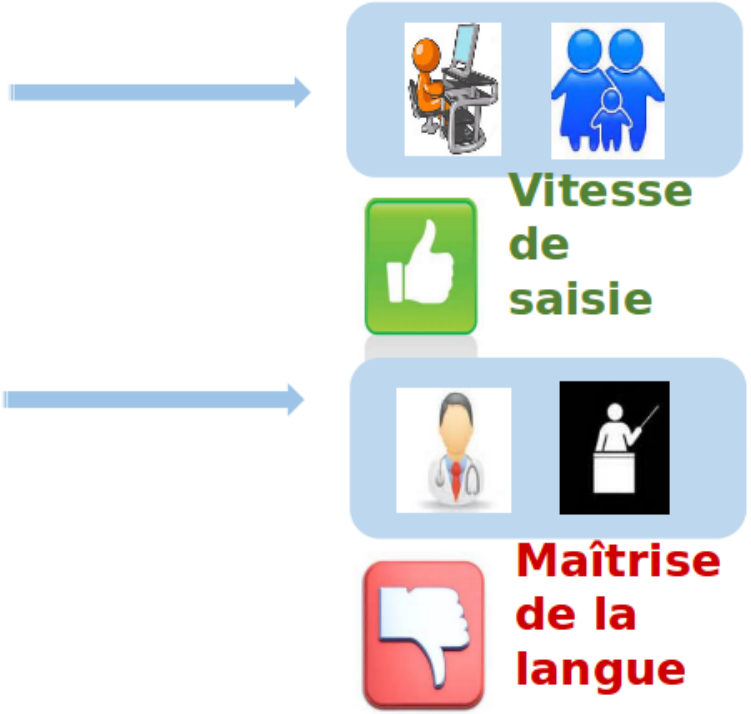
RÉSUMÉ

Nous proposons une réflexion sur les pratiques d'évaluation des systèmes de reconnaissance automatique de la parole (ASR). Après avoir défini la notion de discrimination d'un point de vue légal et la notion d'équité dans les systèmes d'intelligence artificielle, nous nous intéressons aux pratiques actuelles lors des grandes campagnes d'évaluation. Nous observons que la variabilité de la parole et plus particulièrement celle de l'individu n'est pas prise en compte dans les protocoles d'évaluation actuels rendant impossible l'étude de biais potentiels dans les systèmes.

[Garnerin et al., 2020]

Long-term *vs* short-term





[Antoine and Lefeuvre, 2014]

Who is talking, from where?

Practicing ethics in NLP

Helping the thought

Guidelines, guidelines everywhere!

Table 1 Overview of AI ethics guidelines and the different issues they cover

	The European Commission's High-Level Expert Group on Artificial Intelligence	Report on The Future of Artificial Intelligence	Beijing AI Principles	OECD Recommendation of the Council on Artificial Intelligence	The Malicious Use of Artificial Intelligence	AI4People	The Asilomar AI Principles	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	AI Now 2019 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montreal Declaration for Responsible Development of Artificial Intelligence	OpenAI Charter	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version for Public Discussion)	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition)	ITI AI Policy Principles	Microsoft AI principles	DeepMind Ethics & Society Principles	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI	number of mentions
authors	(Pekka et al. 2018)	(Haldane et al. 2016)	(Beijing Academy of Artificial Intelligence 2019)	(Organisation for Economic Co-operation and Development 2019)	(Brundage et al. 2018)	(Floridi et al. 2018)	(Future of Life Institute 2017)	(Crawford et al. 2016)	(Campolo et al. 2017)	(Whittaker et al. 2018)	(Crawford et al. 2019)	(Diakopoulos et al.)	(Abrassart et al. 2018)	(OpenAI 2018)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019)	(Information Technology Industry Council 2017)	(Microsoft Corporation 2019)	(DeepMind)	(Google 2018)	(Cutler et al. 2018)	(Partnership on AI 2018)	
key issue	AI principles of the EU	AI principles of the US	AI principles of China	AI principles of the OECD	analysis of abuse scenarios of AI	meta-analysis about principles for the beneficial use of AI	large collection of different principles	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	principles of the FAT ML community	code of ethics released by the Université de Montréal	several short principles for the ethical use of AI	detailed description of ethical aspects in the context of AI	detailed description of ethical aspects in the context of AI	brief guideline about basic ethical principles	short list of keywords for the ethical use of AI	several short principles for the ethical use of AI	several short principles for the ethical use of AI	IBM's short list of keywords for the ethical use of AI	principles of an association between several industry leaders	
privacy protection	✓		✓	✓	✓			✓	✓	✓	✓		✓		✓	✓				✓	✓	✓	18
fairness, non-discrimination, justice	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓	✓	✓	18
accountability	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓				✓	✓	17
transparency, openness	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓				16
safety, cybersecurity	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	16
common good, sustainability, well-being		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	16
human oversight, control, auditing	✓		✓	✓		✓	✓		✓	✓	✓	✓		✓			✓			✓			12
solidarity, inclusion, social cohesion			✓	✓		✓	✓	✓		✓	✓		✓			✓		✓	✓		✓	✓	11
explainability, interpretability	✓		✓			✓	✓	✓		✓		✓	✓			✓							10
science-policy link		✓	✓		✓	✓	✓	✓		✓				✓		✓			✓				10
legislative framework, legal status of AI systems	✓	✓		✓	✓			✓		✓	✓				✓	✓							10
future of employment/worker rights		✓	✓	✓				✓			✓				✓	✓	✓					✓	9
responsible/intensified research funding		✓		✓		✓	✓	✓							✓		✓		✓				8
public awareness, education about AI and its risks		✓	✓			✓	✓	✓							✓	✓	✓					✓	8
dual-use problem, military, AI arms race		✓			✓		✓	✓	✓	✓	✓		✓		✓	✓				✓			8
field-specific deliberations (health, military, mobility etc.)		✓			✓			✓		✓	✓				✓	✓							8
human autonomy	✓		✓	✓		✓	✓						✓			✓							7
diversity in the field of AI								✓	✓	✓	✓		✓				✓		✓				7
certification for AI products						✓									✓						✓		4
protection of whistleblowers										✓	✓					✓							3
cultural differences in the ethically aligned design of AI systems															✓	✓							2
hidden costs (labeling, clickwork, centered moderation, energy, resources)										✓	✓												2
notes on technical implementations	yes, but very few	none	none	none	yes	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	
proportion of women among authors (f/m)	(6/18)	(2/3)	n.a.	n.a.	(5/21)	(5/8)	n.a.	(4/2)	(1/1)	(6/4)	(12/4)	(1/12)	(8/18)	n.a.	varies in each chapter	varies in each chapter	n.a.	n.a.	n.a.	n.a.	(1/2)	n.a.	(54/77)
length (number of words)	16546	22797	766	3289	34017	3609	646	11532	18273	25759	18970	1358	4756	441	40915	108092	2272	75	417	982	4488	1481	
affiliation (government, industry, science)	government	government	science/gov./ind.	government	science	science	science	science	science	science	science	science	science	non-profit	industry	industry	industry	industry	industry	industry	industry	industry	
number of ethical aspects	9	12	13	12	8	14	13	13	9	12	13	5	11	4	16	18	9	6	6	6	6	8	

[Hagendorff, 2020]

Inspiring sources

- High-Level Expert Group on AI Ethics guidelines for trustworthy AI
- CERNA on Research ethics in ML
- Unesco about Gender divides in digital skills
- CNIL about voice assistants

(Some) systemic approaches to the problems faced in NLP

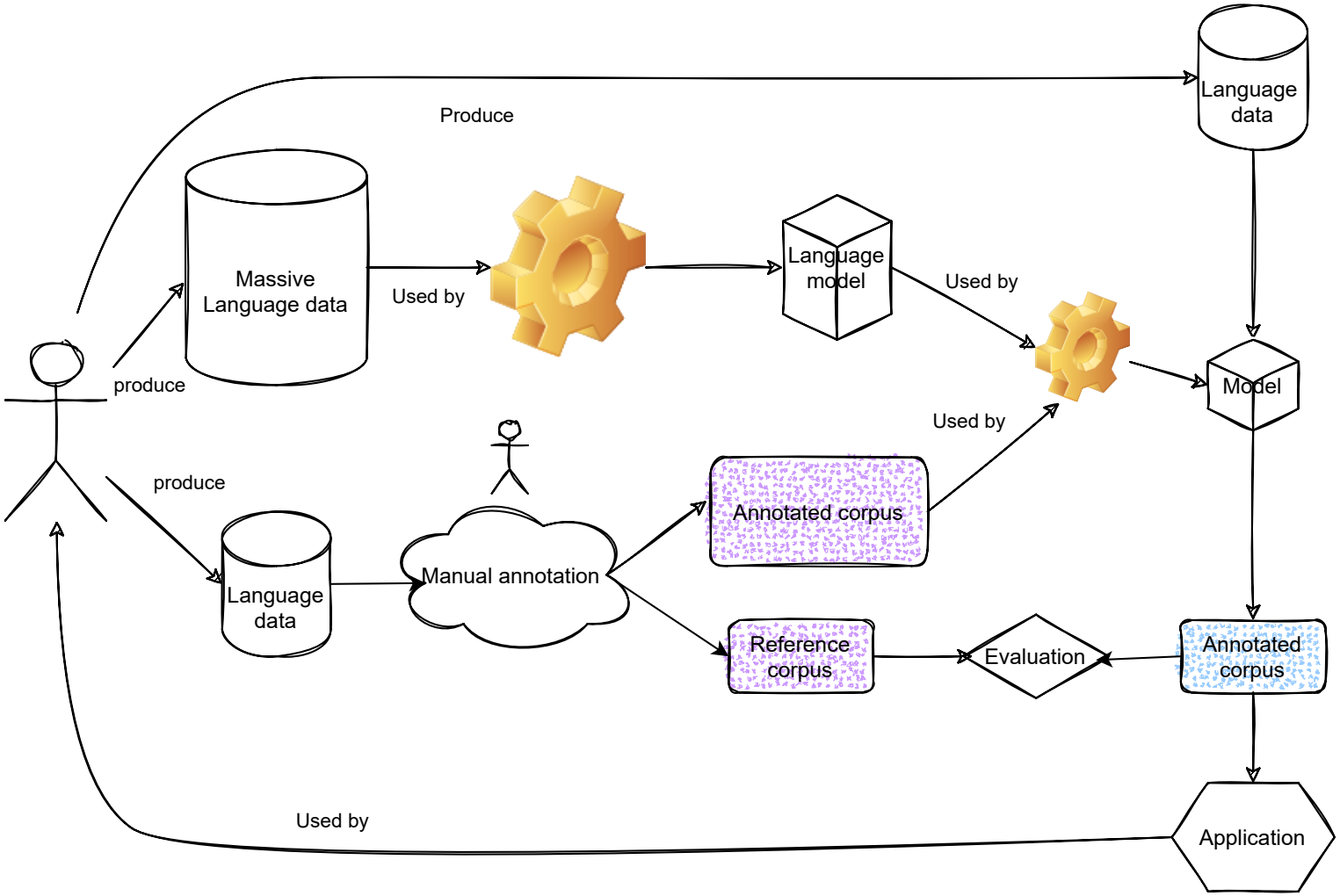
- [Lefeuvre et al., 2015] (in French): a consequentialist grid for an ethical assessment of researches and applications
- [Fort and Amblard, 2018] (in French): a deontological, systemic view on ethics in NLP
- [Bender et al., 2021]: the dangers of large language models (impact on people a posteriori)

Thank you!


We Can Do It!
POST FEB. 15 TO FEB. 28
WAR PRODUCTION CO-ORDINATING COMMITTEE

An overview of (the vast majority of) present NLP

since 2018 (and evolving rapidly) – a draft



Antoine, J.-Y. and Lefeuvre, A. (2014).
Pour une réflexion éthique sur les conséquences de l'usage des NTIC : le cas des aides techniques (à composante langagière ou non) aux personnes handicapées.
In *Actes de la journée ATALA Éthique et TAL*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021).
On the dangers of stochastic parrots: Can language models be too big?
In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Bender, E. M. and Koller, A. (2020).
Climbing towards NLU: On meaning, form, and understanding in the age of data.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Bird, S. (2020).
Decolonising speech and language technology.
In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Blodgett, S. L., Barocas, S., DaumII, H., and Wallach, H. (2020).
Language (technology) is power: A critical survey of "bias" in nlp.
In *ACL*.

Fort, K., Adda, G., and Cohen, K. B. (2011).
Amazon Mechanical Turk: Gold mine or coal mine?
Computational Linguistics (editorial), 37(2):413–420.

Fort, K. and Amblard, M. (2018).
Éthique et traitement automatique des langues.
In *Journée éthique et intelligence artificielle*, Nancy, France.

Garnerin, M., Rossato, S., and Besacier, L. (2020).
Pratiques d'évaluation en ASR et biais de performance.
In Adda, G., Amblard, M., and Fort, K., editors, *2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 1–9, Nancy, France. ATALA.

Hagendorff, T. (2020).
The ethics of ai ethics: An evaluation of guidelines.
Minds & Machines, 30:99–120.

Hovy, D. and Spruit, S. L. (2016).
The social impact of natural language processing.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Lefeuvre, A., Antoine, J.-Y., and Allegre, W. (2015).

Ethique conséquentialiste et traitement automatique des langues : une typologie de facteurs de risques adaptée aux technologies langagières.
In *Atelier Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), conférence TALN'2015*, Actes de la 1e Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), Caen (France), pages 53–66, Caen, France.

Strubell, E., Ganesh, A., and McCallum, A. (2019).
Energy and policy considerations for deep learning in NLP.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017).
Men also like shopping: Reducing gender bias amplification using corpus-level constraints.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.