Flux d'informations dans les systèmes encodeur-décodeur

Application à l'explication des biais de genre dans les systèmes de traduction automatique

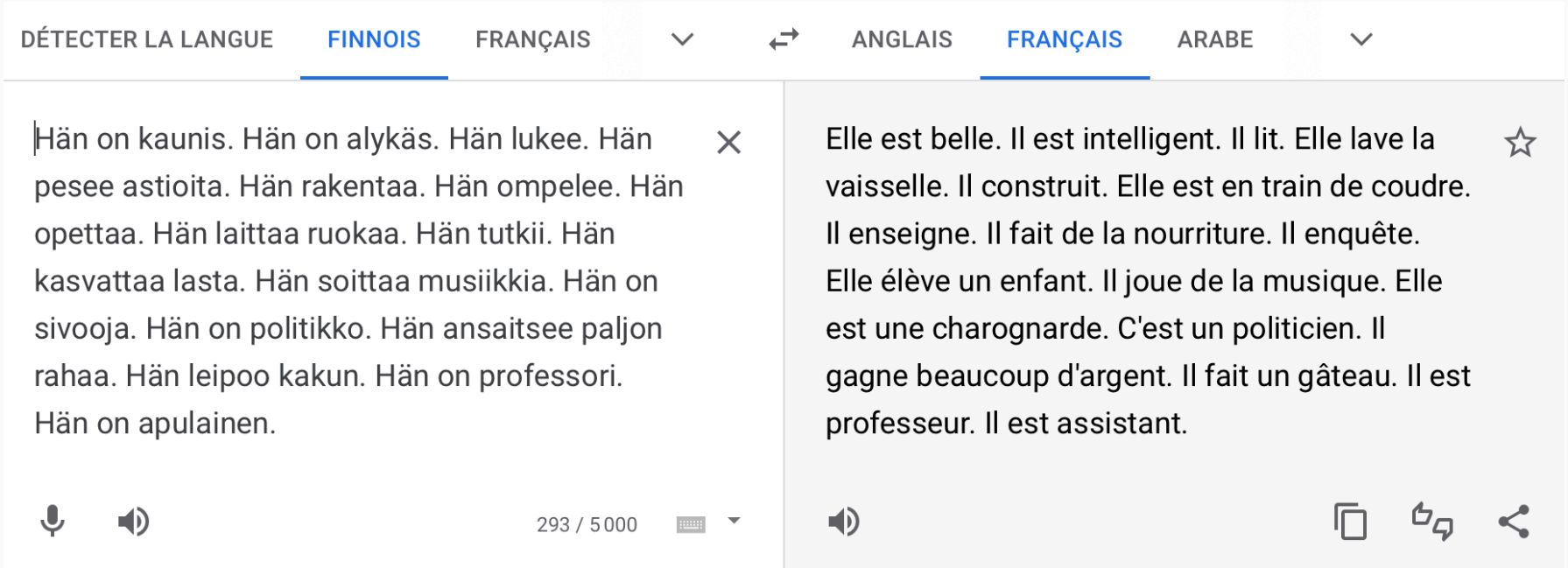
Lichao Zhu[†] Guillaume Wisniewski[†] Nicolas Ballier[‡] François Yvon[♣]

TALN-RÉCITAL 2022 – Atelier *TAL et Humanités Numériques* – 27 juin 2022

[†] Université Paris Cité, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France
[‡] Université Paris Cité, CLILLAC-ARP, F-75013 Paris, France
[♣] Université Paris-Saclay, CNRS, LISN

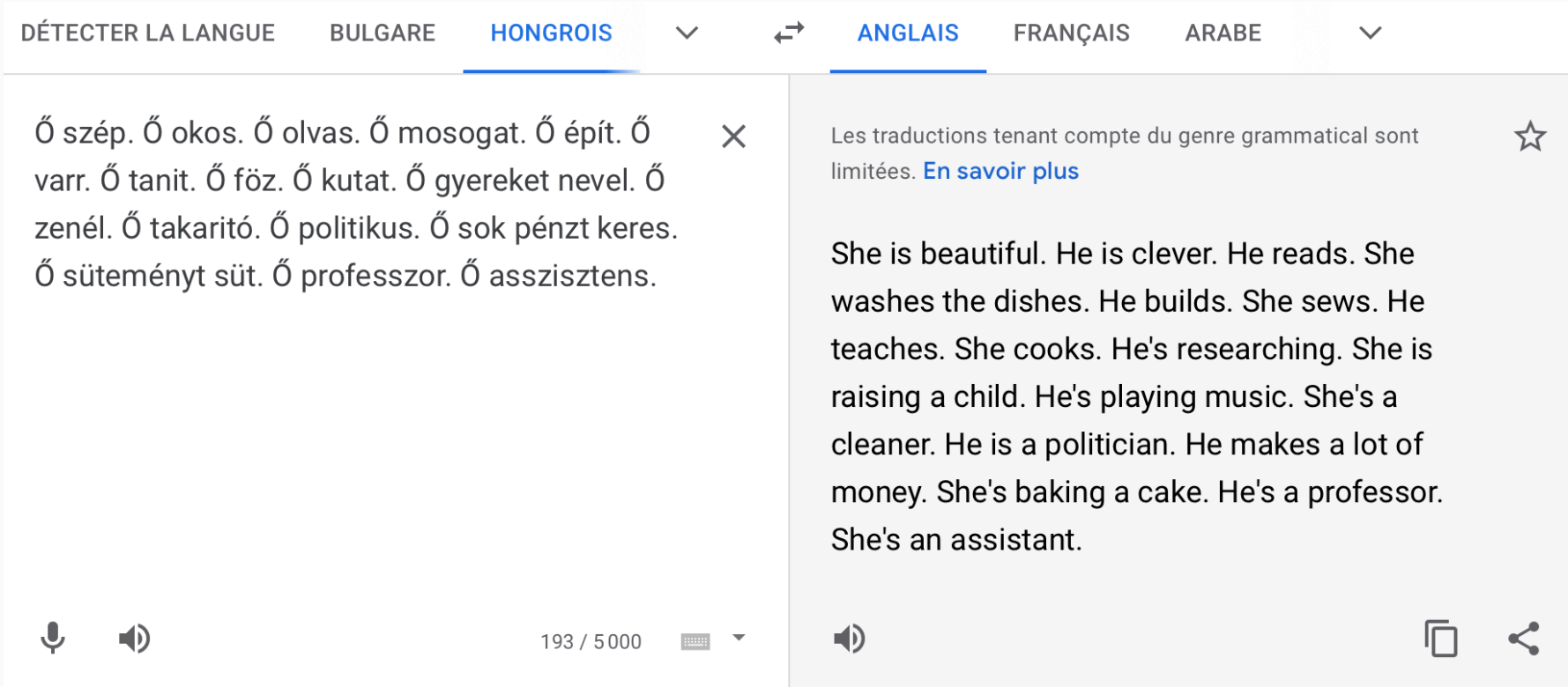
Traduction & biais de genre

En finnois : pronom (**hän**) non genré, mais :



Traduction & biais de genre

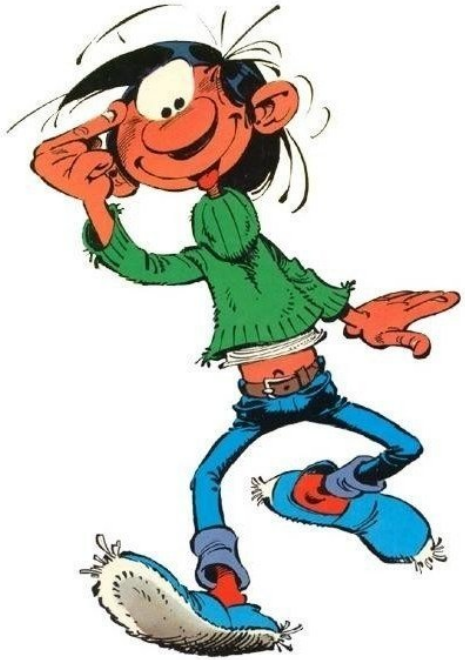
En hongrois : pronom (ő) non genré, mais :



Traduction & biais de genre

- biais de genre dans d'autres langues aussi : swahili, bulgare, hindi, ...
- question de *fairness*
  - ↪ "unfairness" du système de traduction : manque de généralité, erreurs d'output liées au genre, etc. (Stanczak and Augenstein 2021; Savoldi et al. 2021)
  - ↪ conséquences d'un système biaisé : "**Representational harms**" (R) et "**Allocational harms**" (A) (Crawford 2017)

Objectif de ce travail



caractériser les flux d'informations relatifs au genre entre l'encodeur et le décodeur du système de traduction neuronale

Comment le décodeur fait son choix dans la traduction du genre grammatical ?

Plan

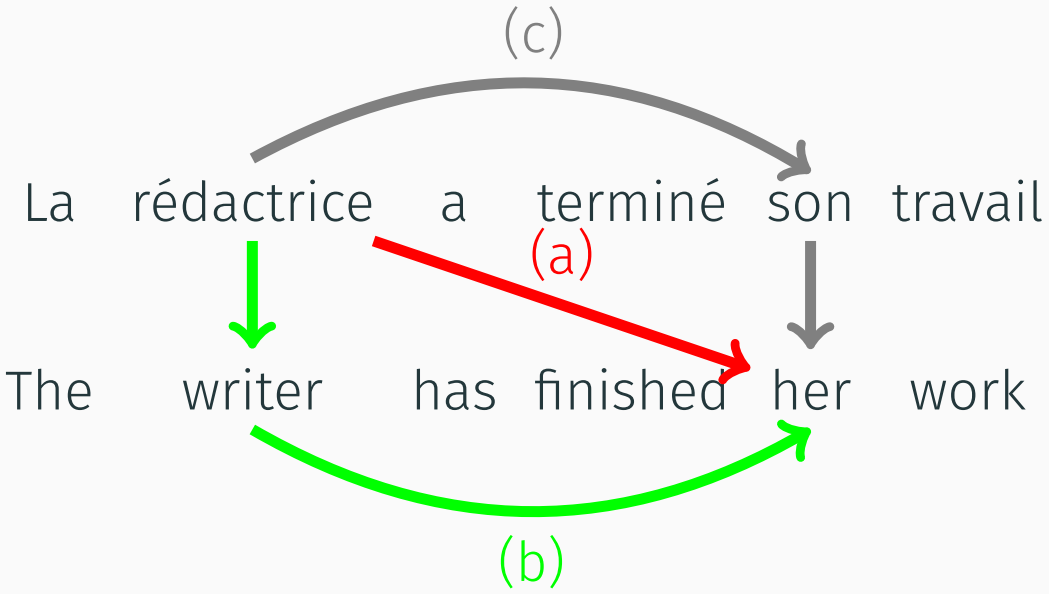
Trois séries d'expérience pour mieux comprendre les flux d'information :

1. Série 1 : Analyses de la propagation de l'information de genre
2. Série 2 : Caractérisation du biais à travers la comparaison entre le TM et le LM
3. Série 3 : Impact sur la prédiction du genre et l'apprentissage

Part I

Analyses de la propagation de l'information de genre

Objectifs



Trois manières de transférer l'information de genre entre l'anglais et le français selon nos connaissances linguistiques

↪ Lesquelles sont utilisées ? Ou aucune des trois n'est utilisée ?

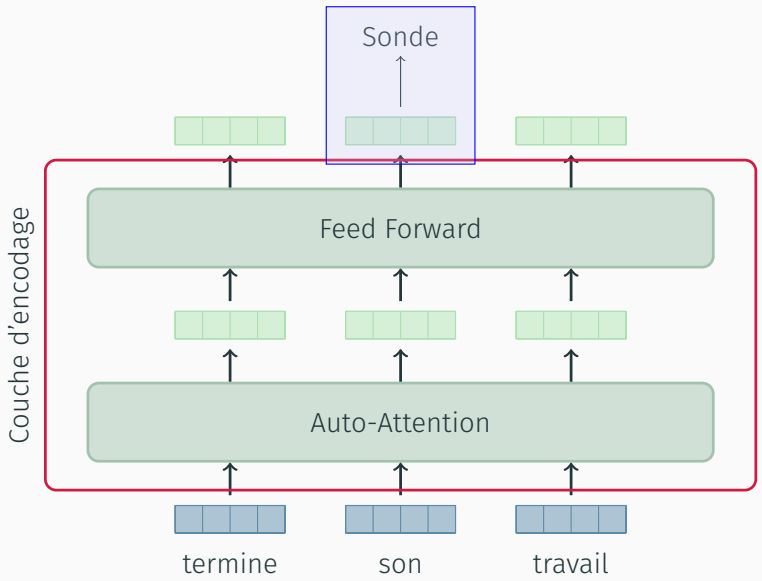
Jeu de données

- constitution d'une liste de noms de métier (> 3 300 noms)
  - ↪ nom de métier dans sa forme masculine et féminine : *infirmier*/*infirmière*
  - ↪ traduction en anglais : *nurse*/*nurse*
- création d'un corpus parallèle

(le|la|l')	X	a	terminé	son	travail	.
the	X	has	finished	(his|her)	job	.

- pour chaque métier : une version « féminine » de la phrase, une version « masculine »

Sonde linguistique : principe



- encodeur : construire une représentation contextualisée des tokens sources
- sonde : utiliser un classifieur supervisé pour prédire une propriété linguistique à partir de la représentation d'un mot

↪ est-ce que la représentation encode une propriété linguistique donnée ?

Sonde linguistique : méthode

- régression logistique : vecteur représentant chaque token de la phrase
- régularisation ℓ_1 (plus de caractéristiques que d'exemples)

Sonde linguistique : résultats

couche	encodeur					
	a	terminé	son	travail	.	*eos*
1	80, 4	75, 1	80, 6	76, 4	59, 5	73, 3
	±1, 1	±0, 3	±0, 3	±0, 6	±1, 0	±1, 0
6	91, 0	89, 3	90, 0	86, 0	86, 4	85, 1
	±0, 6	±0, 2	±0, 2	±1, 0	±1, 1	±0, 8

$\hookrightarrow$ correction de prédiction en hausse de couche en couche

$\Longrightarrow$ le système prédit avec confiance le genre de *son*

$\Longrightarrow$ il existe un flux d'information entre *son* et le nom de métier

Sonde linguistiques : interventions et résultats

	couche	encodeur a	terminé	son	travail	.	*eos*
Affaiblissement							
chaque surveillant a terminé son travail.	1	73, 1	73, 6	65.7	63, 5	53, 9	56, 7
	6	71, 0	71, 4	70, 4	68, 2	71, 2	69, 7
Renforcement							
le surveillant *français* a terminé son travail.	1	99, 9	98, 5	95, 0	80, 6	62, 0	80, 4
	6	100, 0	99, 7	99, 7	98, 9	98, 8	96, 9
...							
...	...	...	...	...	...	...	...

↪ nous avons effectué toute une série d'interventions syntaxiques sur l'information de genre à partir du patron initial (Wisniewski et al. 2021)

↪ le système prédit toujours avec un haut de gré de confiance dans l'ensemble

Part II

Caractérisation du biais à travers la comparaison entre le TM et le LM

Modèle de traduction *vs.* Modèle de langue

Modèle de langue

- prédiction du token cible i conditionnée par le préfixe cible déjà généré i : $p(t_i|t_{<i})$

Modèle de traduction

- modèle de langue conditionnel : $p(t_i|t_{<i}, \mathbf{s})$
  - la phrase source $\mathbf{s}$
  - le préfixe de i : $t_{<i} = t_1, ..., t_{i-1}$

Principe

Modèle de langue

- ✅ la cible

Modèle de traduction

- ✅ la cible
- ☺ la source

↪ La comparaison des deux permet de quantifier et caractériser l'information véhiculée de la source à la cible

Résultats

- **FR** le président Barack Obama a pris note que *[DET] [N]* a mené à bien son travail.
- **EN** President Barack Obama took note that the *[N]* has carried out *[PRO]* work.

	entropie				rang			
	moyenne		médiane		moyenne		médiane	
	LM	TM	LM	TM	LM	TM	LM	TM
_president	5,65	1,12	5,65	1,08	172	1	172	1
_barack	3,70	0,78	3,70	0,79	7	1	7	1
_obama	0,01	1,47	0,01	1,47	1	1	1	1
_took	4,30	2,86	4,30	2,86	34	3	34	3
_note	3,46	0,90	3,46	0,90	23	1	23	1
_that	0,85	1,89	0,85	1,89	5	1,03	5	1
_the	4,46	2,27	4,46	2,10	1	1,08	1	1
noun@0	5,84	3,89	5,84	3,81	9004,92	863,44	7882,50	1
noun@1	3,76	3,03	3,93	2,92	1188,73	487,48	79	1
noun@2	3,77	2,53	4,08	2,09	1039,03	403,31	19	1
noun≥3	4,25	2,68	4,33	2,09	1234,09	477,71	80	1
_has	5,08	3,78	5,26	3,71	20,92	10,07	6	1
_carried	5,42	3,33	5,53	3,31	196,18	2,98	177,50	3
_out	1,79	2,10	1,53	2,11	1,06	1	1	1
_her	4,00	2,18	3,99	2,17	62,62	3,03	50	3
_his	3,99	1,78	3,99	1,72	3,95	1,03	3	1
_work	5,85	3,23	5,91	3,22	4,57	1	3	1
_.	3,43	3,10	3,42	3,14	214,97	1,07	201	1

Part III

Impact sur la prédiction du genre et l'apprentissage

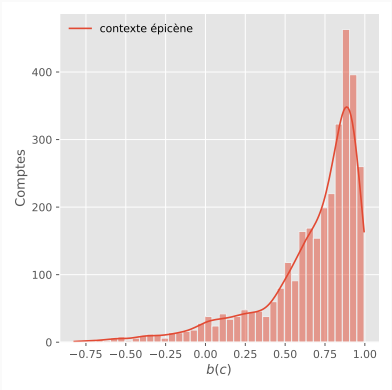
Impact sur la prédiction du genre : score $b(c)$

On utilise le principe (LM *vs.* TM) pour étudier le genre :

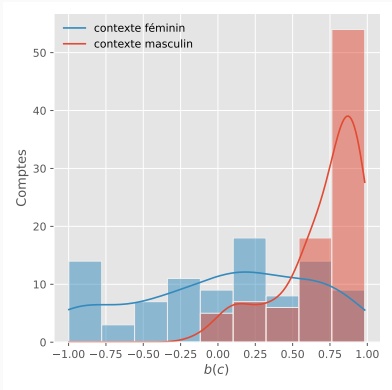
$$b(c) = 1 - \frac{2 \times p(\texttt{her}|c)}{p(\texttt{his}|c) + p(\texttt{her}|c)}$$

- $\hookrightarrow$ c : **contexte** que ce soit TM ou LM
- $\hookrightarrow$ plus $b(c)$ est proche de -1, plus le modèle préfère ***her***
- $\hookrightarrow$ plus $b(c)$ est proche de 1, plus le modèle préfère ***his***

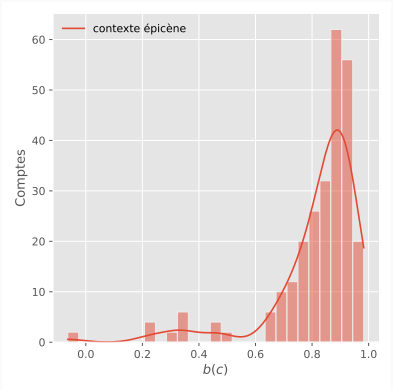
Prédictions du genre : TM *vs.* LM



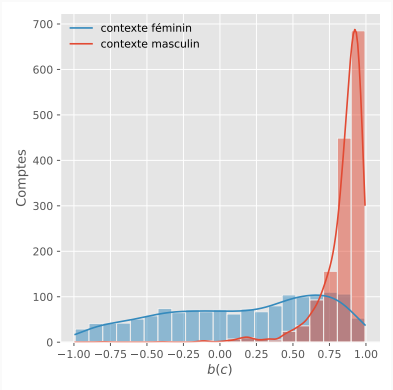
(a) LM
Genre non marqué



(b) LM
Genre marqué



(c) TM
Genre non marqué



(d) TM
Genre marqué

↪ le TM réussit un peu mieux à rééquilibrer les probabilités des deux genres que le LM

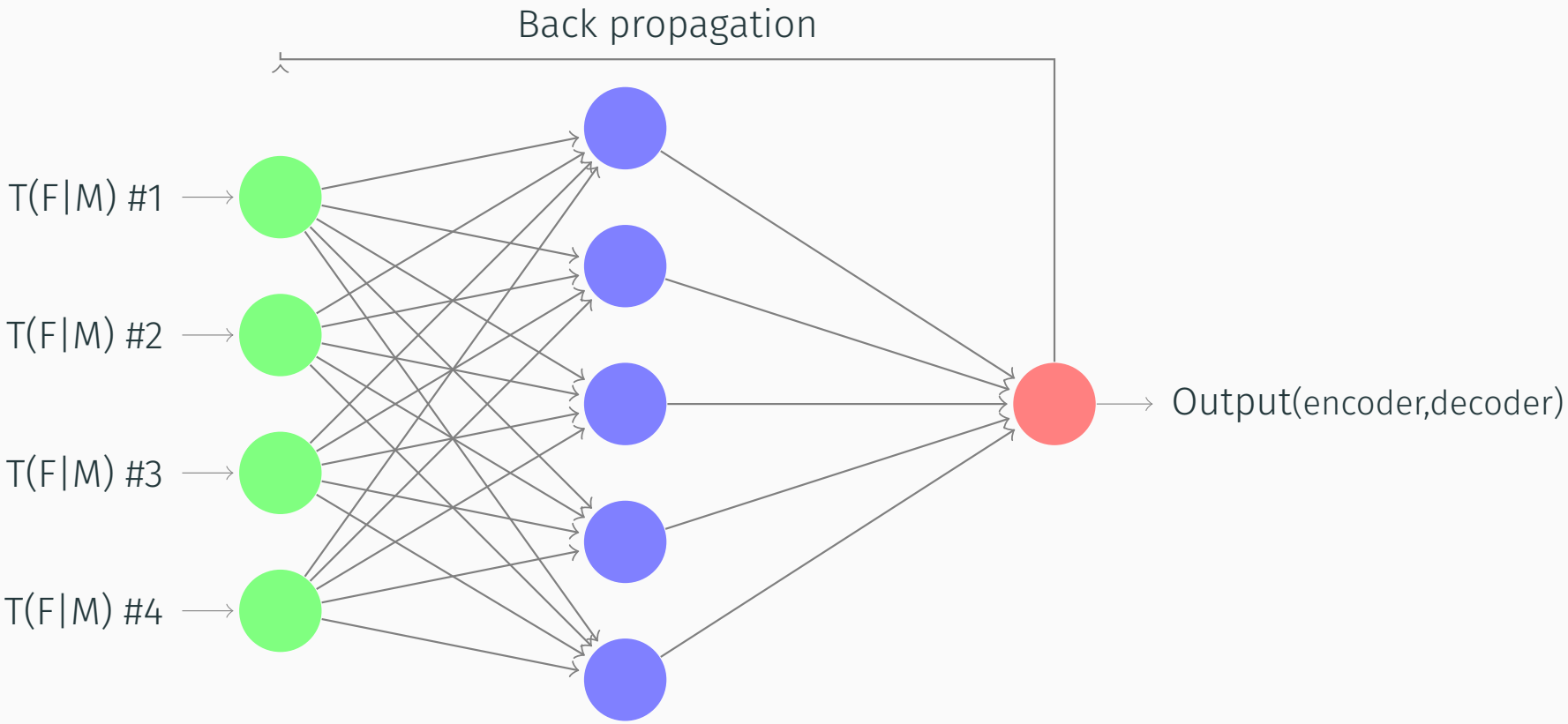
$\Longrightarrow$ le modèle de traduction est capable d'apprendre l'information du féminin de la source lorsque cela est nécessaire

Corpus d'apprentissage contrôlé

- ! ce que l'on sait : la prédiction de *her* s'appuie plus sur le TM, par rapport à celle de *his*
- ? ce que l'on veut savoir : ce comportement peut-il aussi impacter l'apprentissage contrôlé genré ?

Méthode : apprentissage par rétro-propagation

- réaliser une étape d'apprentissage avec, respectivement, les masculins et les féminins
- extraire les gradients accumulés de l'encodeur et du décodeur à l'issue de chaque étape



Résultats

	couche	$\frac{\nabla \text{param}_{\text{masc}}}{\nabla \text{param}_{\text{fémi}}}$
décodeur	0	0.719517
	1	0.756060
	2	0.758951
	3	0.720758
	4	0.780754
	5	0.950173
encodeur	0	0.652739
	1	0.649395
	2	0.713145
	3	0.661217
	4	0.729006
	5	0.770339

↪ les gradients accumulés sur les exemples féminins sont systématiquement plus grands que ceux des masculins

↪ le système cherche des informations dans la source pour ”corriger” les prédictions erronées des exemples féminins

Conclusion

Analyse des informations utilisées pour prédire le genre

↪ sonde linguistique

↪ intervention sur la représentation

↪ analyse causale

Enjeu sociétal

↪ la question de *fairness* mise en lumière

↪ IA au service ou complice ??

Merci de votre attention