



Myriadisation et éthique pour le traitement automatique des langues

Karën Fort

► To cite this version:

Karën Fort. Myriadisation et éthique pour le traitement automatique des langues. Traitement du texte et du document. ED n°77: Informatique - Automatique - Électronique - Électrotechnique - Mathématiques de Lorraine (IAEM-Lorraine), 2022. tel-03873000

HAL Id: tel-03873000

<https://theses.hal.science/tel-03873000>

Submitted on 28 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

Myriadisation et éthique pour le traitement automatique des langues

Mémoire déposé, présenté et soutenu publiquement le 23 novembre 2022

pour l'obtention de l'

Habilitation à Diriger des Recherches de l'Université de Lorraine

(spécialité informatique)

par

Karèn Fort

Composition du jury

Rapporteurs :	Jean-Yves Antoine, Iris Eshkol, Dirk Hovy,	Professeur, Université de Tours Professeure, Université Paris Nanterre Associate Professor, Bocconi University
Examineurs :	Philippe Blache, Armelle Brun, Massimo Poesio, Marta Severo,	Directeur de recherche, CNRS Professeure, Université de Lorraine Professeur, Université Queen Mary, Londres Professeure, Université Paris Nanterre
Marraine :	Claire Gardent,	Directrice de recherche, CNRS

Laboratoire Lorrain de Recherche en Informatique et ses Applications – UMR
7503

à ma grand-mère, Suzanne Rojat,
à ma mère, Michèle Gabert,
sans qui je ne serais pas ce que je suis
à Michel, sans qui je n'aurais jamais su qui je voulais être
pour Léonard, sois ce que tu veux

Remerciements

Je tiens à remercier ici chaleureusement Claire Gardent, qui a accepté d'être ma marraine pour cette habilitation à diriger les recherches et qui m'a soutenue pendant la rédaction, m'aidant à vaincre mes doutes et à avancer.

Je remercie également mes rapporteurs, Jean-Yves Antoine, Iris Eshkol et Dirk Hovy, qui ont pris de leur temps pour me lire et évaluer mon travail. Merci aux autres membres du jury, Philippe Blache, Armelle Brun, Massimo Poesio et Martha Severo, d'avoir accepté d'en faire partie et d'avoir eux aussi donné de leur temps pour cela. Ce travail d'évaluation n'est pas toujours reconnu à sa juste valeur alors qu'il est fondamental.

Cette HDR n'aurait pas vu le jour sans le soutien de mon équipe de recherche, Sémagramme, et plus particulièrement celui de Philippe de Groote, son responsable et de Bruno Guillaume, mon co-bureau et ami. Maxime Amblard m'a fait rire dans les moments tragiques, ce qui n'a pas de prix. Gabriel Sauger, Vincent Tourneur, Valentin Richard et Heesoo Choi m'ont fait jouer au ping-pong quand plus rien n'allait. La « Team LREC » m'a redonné le goût des autres et des diners trop longs. Merci à Priyansh Trivedi, Siyana Pavlova et Fanny Ducel pour leurs rires et leur grand cœur. Merci à Michel Musiol pour ses grandes oreilles.

Si Sémagramme m'aide à garder le cap, Gaël Lejeune me permet de tenir au quotidien, par son amour des autres, son intelligence et son humour. Je lui dois beaucoup.

Alice Millour, ma doctorée, m'a encouragée tout au long de la rédaction et ses remarques sur mon tapuscrit m'ont été très profitables. Merci pour ce travail. Surtout, merci de me faire rire aux larmes.

Ma directrice de thèse, Adeline Nazarenko, reste pour moi une référence dans notre métier et je continue à apprendre beaucoup à son contact. Nos coups de fil me font grandir et ses retours sur mes travaux me sont précieux.

Je voudrais enfin remercier tous les collègues avec lesquels j'ai eu le plaisir de travailler de ces dix dernières années, en particulier Aurélie Névéol et Marc Anderson, avec qui j'apprends beaucoup. Merci à Didier Ozil de m'avoir descendue du piédestal et de m'avoir écoutée.

Merci à mes étudiants de me pousser à me remettre en questions régulièrement et de m'envoyer, parfois, des messages gentils. Je tiens à mentionner ici ceux avec qui j'ai pu collaborer en recherche, notamment dans le cadre de leur mémoire de M1 (et pour certains beaucoup plus) : Ange Richard, Yann-Alan Pilatte, Diego

REMERCIEMENTS

Alves, Harmonie Begue, Rafael Araujo, Nikola Lackovic, Julien Bezançon, Alexane Jouglar, Heesoo Choi, Nicolas Hiebel et Fanny Ducel.

Enfin, merci à mon compagnon, Michel Ancé, et à mon fils, Léonard, de me supporter et de me soutenir, de m'aimer, comme je suis.

Résumé

Le traitement automatique des langues (TAL) a subi deux révolutions ces dix dernières années : le raccourcissement extrême de la distance entre les productions de la recherche et l'utilisateur final et l'avènement de l'apprentissage profond (deep learning). En conséquence, les besoins en données ont explosé en parallèle des questions éthiques. Cette habilitation à diriger des recherches présente les travaux que j'ai menés dans le domaine de la production d'annotations manuelles pour le TAL par myriadisation (crowdsourcing), en particulier par le jeu (games with a purpose), et dans celui de l'éthique pour le TAL. J'y redéfini la myriadisation et les sciences participatives en général et je présente en détail les jeux ayant un but, leurs atouts et leurs limites. Je m'attarde plus particulièrement sur ZombiLingo, qui a servi à collecter des annotations en syntaxe de dépendances pour le français et RigorMortis, un jeu d'annotation d'unités polylexicales. Je me concentre dans une dernière partie sur l'éthique pour le TAL, un sous-domaine qui n'a véritablement été reconnu qu'à partir de 2016 et dont j'ai été précurseure. Je reviens sur son historique, son évolution récente et présente mes travaux, menés dans une optique plus déontologiste que conséquentialiste, permettant d'avoir une vision systémique du TAL et des problèmes éthiques qu'il pose.

Mots clés :

Traitement automatique des langues, myriadisation, jeux ayant un but, éthique.

Abstract

In the past ten years, Natural Language Processing (NLP) has undergone two revolutions : the extreme shortening of the distance between research outputs and the end user and the advent of deep learning. As a result, data needs have exploded alongside ethical issues. This "habilitation à diriger des recherches" presents the work I have carried out in the field of the production of manual annotations for NLP by crowdsourcing, in particular using games with a purpose, and in that of ethics for the TAL. I redefine crowdsourcing and citizen science in general and I present in detail the games with a purpose, their strengths and their limits. I focus more particularly on ZombiLingo, which was used to collect dependency syntax annotations for French, and RigorMortis, a game aiming at collecting multiword expressions (MWE). In the last part, I focus on ethics for NLP, a sub-field that was only truly recognized from 2016 and of which I was a forerunner. I return to its history, its recent evolution and present my work, carried out in a more deontological than consequentialist perspective, allowing to have a systemic vision of NLP and the ethical problems it poses.

Key words :

Natural Language Processing, crowdsourcing, games with a purpose, ethics.

Table des matières

Introduction : faire chemin	1
1 La myriadisation : miroir grossissant sur l’annotation manuelle de corpus	5
1.1 Myriadisation et <i>Crowdsourcing</i> : du jeu dans les termes	5
1.2 Un continuum de pratiques	6
1.2.1 (Encore) une typologie des myriadisations	6
1.2.2 Le travail parcellisé	7
1.2.3 La myriadisation bénévole	8
1.2.4 Les jeux ayant un but	9
1.3 Des mythes à déconstruire	10
1.3.1 Mythe #1 : la myriadisation est un phénomène récent	10
1.3.2 Mythe #2 : la myriadisation implique une foule de participants	11
1.3.3 Mythe #3 : la myriadisation implique des non-experts	12
1.4 Myriadisation et sciences participatives	13
1.4.1 Pour une définition éthique des sciences participatives	14
1.4.2 Les données pas données des sciences participatives	16
1.5 Annotation manuelle de corpus et myriadisation	17
1.5.1 Annoter, c’est quantifier	17
1.5.2 Une activité (insuffisamment) outillée	18
1.5.3 Des annotateurs sous influences	19
1.5.4 Le mythe de l’annotateur expert	21
2 Les jeux ayant un but : un modèle encore incertain	25
2.1 Les différents types de jeux ayant un but	25
2.1.1 Les jeux faisant appel aux connaissances du monde et de la langue des locuteurs	26
2.1.2 Les jeux faisant appel aux connaissances scolaires des locuteurs	28
2.1.3 Les jeux faisant appel aux capacités d’apprentissage des lo- cuteurs	30
2.2 Ludifier la syntaxe en dépendances : <i>ZombiLingo</i>	31
2.2.1 Une tâche complexe	32
2.2.2 Décomplexifier n’est pas simplifier	32

TABLE DES MATIÈRES

2.2.3	Mécanismes de formation et de contrôle	33
2.2.4	Des résultats très encourageants	35
2.2.5	De ZombiLingo à ZombiLUDik	37
2.3	Des jeux aux plateformes (moins) ludifiées	37
2.3.1	Rigor Mortis : annoter les unités polylexicales	38
2.3.2	Bisame et Recettes de grammaire : appliquer la myriadi- sation aux langues non-standardisées	40
2.4	Les incertitudes du modèle	45
2.4.1	Créer un jeu, un savoir-faire complexe	46
2.4.2	Des contraintes des jeux ayant un but	47
2.4.3	Motiver les participants	48
3	L'éthique dans et pour le TAL	55
3.1	Créer le chemin vers la légitimité	55
3.1.1	Des questions invisibles en TAL avant 2011	55
3.1.2	Des espaces pour réfléchir, ensemble	56
3.1.3	D'une dynamique locale à une reconnaissance internationale	58
3.1.4	Des oppositions peu spécifiques au TAL	60
3.2	Une éthique morcelée par la spécialisation	63
3.2.1	Limiter les biais	64
3.2.2	Améliorer la diversité linguistique	66
3.2.3	Documenter les données	69
3.2.4	Limiter l'empreinte carbone	70
3.3	Pour une vision plus systémique de l'éthique dans le TAL	71
3.3.1	Les éthiques philosophiques	72
3.3.2	Des grilles d'analyse conséquentialistes pour le TAL	73
3.3.3	Vers une grille d'analyse systémique pour le TAL	73
4	Perspectives et projet de recherche	79
4.1	Un environnement structurant	79
4.2	Produire (de manière éthique) des ressources langagières pour le TAL	80
4.2.1	Des ressources sémantiques libres pour le français	80
4.2.2	Des ressources synthétiques pour préserver le secret médical	82
4.2.3	Des ressources pour évaluer les biais dans les modèles	82
4.3	Mener la réflexion éthique en TAL et en IA	83
4.3.1	Favoriser la réflexion sur la recherche en TAL	83
4.3.2	Participer à la dynamique « Éthique et IA » du LORIA	86
4.4	Croiser les chemins	87

Table des figures

1	Le TAL aujourd’hui, vu de l’annotation manuelle.	2
1.1	Sens 4 de la définition de <i>crowd</i> dans le <i>Merriam Webster</i>	6
1.2	Une typologie de la myriadisation selon deux axes : la rémunération et la transparence de la tâche (le participant est-il conscient de ce qu’il construit ?).	7
1.3	Nombre de joueurs sur <i>Phrase Detectives</i> selon le nombre de points gagnés dans le jeu (fév. 2011 - fév. 2012) (Chamberlain et al., 2013).	12
1.4	Commentaire du joueur Justin dans <i>ZombiLingo</i> à propos de la différence entre le passif et le passé composé dans l’annotation en question.	13
1.5	Commentaire du joueur Zeltron dans <i>ZombiLingo</i> à propos de « fini », qui n’est pas un verbe au passif dans l’annotation en question.	13
1.6	Participation à des groupes de travail sur les sciences participatives (en bleu, CNRS, en vert, Sorbonne Université, en violet, Ministère de la culture).	14
1.7	Page de sélection des relations à annoter dans <i>ZombiLingo</i>	20
1.8	Extrait du corpus de presse ancienne de la campagne Quæro, à annoter en entités nommées (Rosset et al., 2012).	21
1.9	Phrase d’EMEA à annoter par étapes dans <i>ZombiLingo</i> : ici, la bonne réponse est « perfusion ».	22
2.1	Figure 1 de von Ahn and Dabbish (2004) : deux participants jouant sur <i>ESP Game</i> et ayant trouvé la même étiquette (« purse »).	26
2.2	Interface de <i>JeuxDeMots</i> : le joueur doit trouver des idées (termes) associées à « qualité ».	27
2.3	Interface de <i>Phrase Detectives</i> : le joueur doit trouver un antécédent de « a theme » (s’il existe).	28
2.4	Interface de <i>Phrase Detectives</i> : le joueur doit (in)valider l’annotation d’un autre joueur entre « Knitta taggers » et « The group ».	29
2.5	Formation par étapes dans <i>FoldIt</i> (à gauche) et interface du jeu (Figure 4 de (Cooper et al., 2010)).	30
2.6	Progression des trois joueurs de <i>FoldIt</i> vers la solution (Figure 2 de (Khatib et al., 2011)).	31

TABLE DES FIGURES

2.7	Exemple d'annotation en syntaxe de dépendances selon le schéma d'annotation du corpus Sequoia.	32
2.8	Interface de sélection de relations à jouer dans ZombiLingo	33
2.9	Interface principale de jeu dans ZombiLingo : l'élément en surbrillance est la tête de la relation, le joueur doit trouver le dépendant (ici, « à »).	33
2.10	ZombiLingo : formation obligatoire avant de pouvoir jouer la relation (ici, modificateur).	34
2.11	ZombiLingo : retour au joueur dans le cas d'une erreur lors de la formation, l'erreur est signalée et la solution est indiquée.	34
2.12	Organisation de l'évaluation de ZombiLingo	35
2.13	F-mesures pour les deux analyseurs syntaxiques utilisés pour la pré-annotation et le jeu, par relation (celles avec une densité supérieure à 1 sont sur la gauche, les autres à droite séparées par une espace).	36
2.14	Interface principale de Rigor Mortis	38
2.15	Phase deux de Rigor Mortis : formation des participants.	39
2.16	Figure 8 de (Fort et al., 2020) : précision, rappel et F-mesure en fonction du seuil d'accord.	41
2.17	Annotation directe.	42
2.18	Bandeau d'accueil de Recettes de grammaire pour l'alsacien.	43
2.19	Badges que les participants peuvent obtenir dans Recettes de grammaire pour l'alsacien, en fonction de leur activité sur les différentes tâches.	44
2.20	Ajout de variantes dans Recettes de grammaire (alsacien).	44
2.21	Typologie des joueurs selon (Bartle, 1996).	47
2.22	Katana et Grand Guru : les interactions entre le jeune joueur et son aîné permettent à la fois de produire des données langagières et de favoriser la transmission de la langue entre générations.	50
2.23	Katana et Grand Guru : le mot irlandais pour arbre (« crann ») permet de faire reverdir les arbres, mais en cas d'erreur les arbres restent morts.	51
2.24	Katana et Grand Guru : scénario du premier niveau.	52
3.1	Activités autour de l'éthique dans le TAL sur la dernière décennie.	57
3.2	Les modèles de langue au centre du TAL actuel.	64
3.3	Les cinq sources de biais dans le TAL actuel selon (Hovy and Prabhu-moye, 2021).	65
3.4	Proportion d'articles appliquant la règle de Bender pour chaque édition des trois conférences (ACL en noir, LREC en rouge et TALN en bleu).	68

TABLE DES FIGURES

3.5	Évolution de la production des modèles de langues entraînés ces dernières années en taille du jeu de données d’entraînement et en nombre de paramètres, schéma issu d’une présentation de l’article (Bender et al., 2021).	71
3.6	Environnement de production de la recherche (rose et blanc), acteurs (bleu), données (vert).	74
3.7	Les flèches violettes montrent les flux de données et illustrent la difficulté du consentement et de la traçabilité.	76
4.1	Projets en cours (en vert les actions COST, en rouge les projets européens, en bleu les projets ANR).	80

TABLE DES FIGURES

[...] parfois, je crois qu'on peut vraiment bâtir, non pas une œuvre, mais un ouvrage, à partir de ses défauts. Et je crois que la singularité de certains grands artistes, c'est justement leur incapacité à faire certaines choses, qui les amène à en faire d'autres. Je crois plus dans les failles des êtres humains que dans leurs certitudes.

Angelin Preljocaj, L'éloge du déséquilibre,
Hors Champs, France Culture, 2013

Introduction : faire chemin

Depuis que j'ai commencé ce travail d'écriture me reviennent en boucle les vers d'Antonio Machado ([Machado, 1912](#)) :

« Caminante, no hay camino, se hace camino al andar. »
[Voyageur, il n'y a pas de chemin, le chemin se fait en marchant]

En effet, dans le contexte de recherche qui est le nôtre aujourd'hui, entre modes, recherches de financements presque constantes et contraintes administratives diverses, il est difficile de tenir un cap et chacun avance comme il le peut, faisant son chemin *malgré* et *avec* les contraintes qui nous sont imposées. Cet exercice de synthèse est donc à la fois une gageure et une opportunité de redonner du contexte et du sens, *a posteriori*, à ce qui a été fait.

Je voudrais insister ici sur le fait que rien de ce que j'ai pu faire n'aurait été possible sans les collègues et les étudiants avec lesquels j'ai eu la chance et le plaisir intense de travailler. Je vais bien entendu les citer tout au long de ce tapuscrit, mais je voulais leur rendre hommage dès ces premières lignes, tant la recherche me semble un exercice bien fade quand il faut le livrer seule.

L'annotation manuelle de corpus face aux bouleversements du TAL

Ce travail de synthèse et de réflexion est mené près de dix ans après ma thèse, soutenue en 2012 ([Fort, 2012](#)). Celle-ci portait sur l'annotation manuelle de corpus pour le traitement automatique des langues (TAL) réalisée de manière traditionnelle, c'est-à-dire avec des annotateurs rémunérés dans le cadre d'un contrat de travail ou d'un stage.

Entre temps, le TAL a connu deux révolutions majeures, étroitement liées : l'avènement et le déploiement massif de l'apprentissage dit « profond » (*deep learning*) et le rapprochement entre la recherche et l'utilisateur final. De ces changements, le deuxième a eu plus d'impact sur mes activités de recherche que le premier, malgré sa prévalence.

Certes, l'annotation manuelle est aujourd'hui moins massivement utilisée, grâce aux modèles de langues de type BERT ([Devlin et al., 2019](#)), mais elle reste fondamentale pour l'ajustement (*fine-tuning*) et l'évaluation des systèmes. Le TAL

d’aujourd’hui, vu de l’annotation manuelle n’a donc que peu évolué, comme le montre la figure 1, dans laquelle la partie sélectionnée en bleu n’a pas fondamentalement changé, même si le contexte autour a été bouleversé.

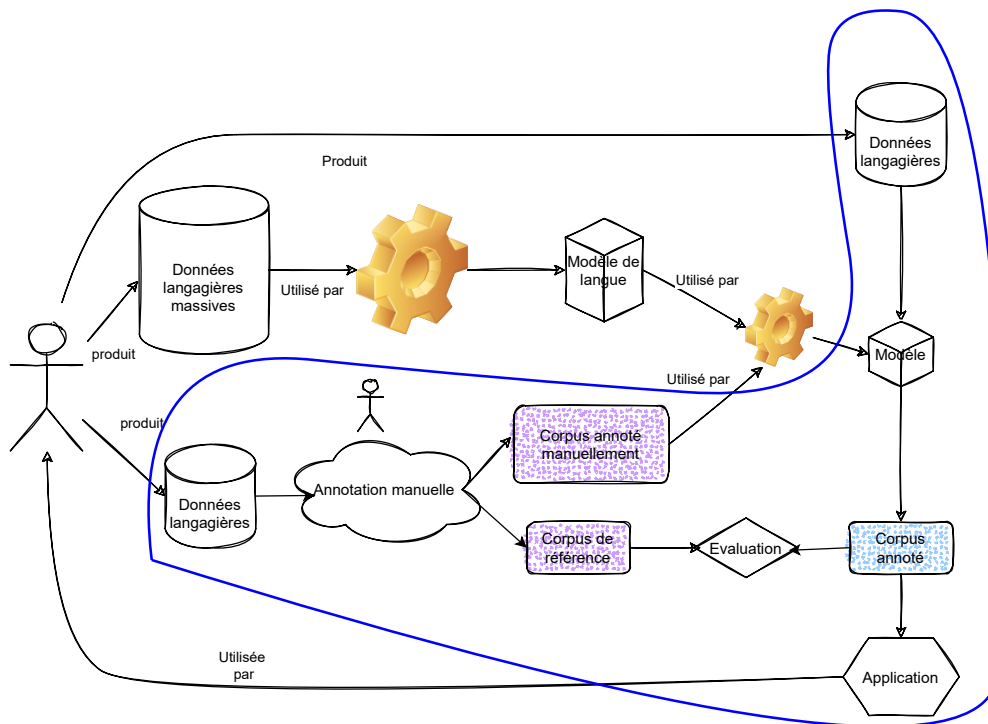


FIGURE 1 – Le TAL aujourd’hui, vu de l’annotation manuelle.

Cela ne signifie pas que je ne m’intéresse pas aux technologies d’apprentissage profond et en particulier aux plongements. J’ai ainsi poussé ma doctorante, Alice Millour, à reproduire les expériences d’étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux (Magistry et al., 2018), ce qui nous a amenées à collaborer avec Pierre Magistry (Inalco) (Millour et al., 2020) et à monter en compétences sur le sujet. De même, j’ai travaillé avec Aurélie Névéal (LISN), Yoann Dupont et Julien Bezançon (Sorbonne Université) sur un corpus de stéréotypes du français, grâce auquel nous avons pu tester les modèles de langue du français (Névéal et al., 2022). Enfin, le doctorant que je co-encadre avec Aurélie Névéal et Olivier Ferret (CEA), Nicolas Hiebel, travaille en ce moment sur **Sentence-BERT** (Reimers and Gurevych, 2019). Simplement, je ne suis, comme beaucoup, qu’une utilisatrice, ces technologies ne sont pas au centre de ma recherche, je ne vais donc pas m’étendre dessus ici.

En revanche, le récent rapprochement entre la recherche et l’utilisateur final a brutalement mis en lumière les problèmes éthiques que pose le TAL, notamment

les biais stéréotypés, et fait émerger le sous-domaine de l'éthique du TAL qui est le mien aujourd'hui.

L'éthique comme fil conducteur

Le tournant qu'ont pris mes travaux après ma thèse a été profondément influencé par les questions éthiques que j'ai commencé à me poser pendant celle-ci.

En effet, j'ai commencé à m'intéresser à la myriadisation pour le TAL par le biais des problèmes éthiques que pose la plateforme de travail parcellisé (*microworking*) **Amazon Mechanical Turk**, utilisée pour réaliser des annotations manuelles à moindre coût (Fort et al., 2011). Cherchant une alternative plus éthique à ce type de plateforme, je me suis intéressée dès 2013 aux jeux ayant un but (*Games with a purpose*, ou GWAP, en anglais), avec Bruno Guillaume, chercheur à l'Inria de Nancy, où je travaillais alors dans le cadre de mon contrat d'ATER.

Après mon recrutement à la Sorbonne en 2014, j'ai participé à plusieurs groupes de travail à Paris sur les sciences participatives, où j'ai pu rencontrer des collègues investis dans des projets très différents, comme les plateformes du Museum national d'histoire naturelle (MNHN) **Vigie Nature**¹, les **Herbonautes**² ou moins orientés sur la production de données, comme le **FabLab** de Sorbonne Université³. Cette ouverture nouvelle m'a aidée à mieux définir les sciences participatives, la myriadisation, et à positionner les jeux ayant un but en leur sein.

En parallèle, je continuais à travailler sur la myriadisation pour le TAL, en particulier avec Alice Millour, que j'ai encadrée, à partir de 2015 en Master puis en doctorat⁴, sur la myriadisation pour les langues peu dotées non standardisées.

Cette exploration m'a permis de revenir sur l'annotation manuelle en y portant un regard nouveau, depuis la myriadisation, qui met au jour les failles de l'annotation traditionnelle.

La myriadisation bénévole comme alternative au microtravail et la création de ressources langagières pour les langues peu dotées sont des sujets étroitement liés à la fois à l'éthique et au TAL.

En parallèle, j'ai mené des actions pour construire des espaces réflexifs collectifs et des outils pour l'éthique du TAL. Cette activité prend actuellement beaucoup d'ampleur dans ma vie d'enseignante-chercheuse, du fait de responsabilités importantes au niveau international et des cours que je donne sur le sujet. Je tiens cependant à conserver une activité soutenue en TAL et en création de ressources langagières, afin d'asseoir la réflexion sur la pratique.

1. Voir : <https://www.vigienature.fr/>.

2. Voir : <http://lesherbonautes.mnhn.fr/>.

3. Voir : <http://fablab.sorbonne-universites.fr/>.

4. Alice Millour a soutenu sa thèse en décembre 2020.

Cet ancrage dans le TAL, allié à une ouverture multidisciplinaire sont caractéristiques de mon travail depuis mes débuts en recherche. Sans doute ai-je besoin de ce va-et-vient constant pour affermir ma pensée.

Cheminer : du théorique, au pratique, vers l'éthique

J'ai choisi d'organiser ce tapuscrit de manière un peu circulaire, afin de situer ce dont je parle précisément avant d'en montrer des instanciations, en terminant par ce qui m'a portée et continue de me porter aujourd'hui.

Le chapitre 1 est ainsi consacré à la myriadisation, présentant un point de vue définitoire et typologique sur celle-ci et montrant ce qu'elle peut apporter à l'annotation traditionnelle. J'y reprends certains éléments développés dans mon livre sur l'annotation collaborative de corpus (Fort, 2016), en les reliant avec davantage de recul à mes travaux de thèse.

Je détaille ensuite dans le chapitre 2 ce que sont les jeux ayant un but et ce qu'ils permettent, en illustrant la réflexion par certaines des expériences que j'ai pu mener sur des jeux ou des interfaces ludifiées. Je montre que créer un jeu est un savoir faire complexe, qui nécessite non seulement de trouver une idée porteuse, mais également de gérer les contraintes inhérentes à la création de données de qualité.

Enfin, je consacre le chapitre 3 à mes travaux concernant l'éthique du TAL. J'y présente les actions que j'ai menées depuis une dizaine d'années pour faire vivre la thématique dans le domaine, ainsi que mes travaux de recherche sur le sujet. J'y développe une vision systémique de l'éthique dans le TAL et plus largement dans l'IA, qui permet de couvrir des thèmes qui ne sont pas véritablement abordés aujourd'hui dans le domaine.

Je conclus en abordant les projets auxquels je participe actuellement et l'orientation que je souhaite donner à mes recherches dans le futur, afin de « faire chemin », malgré tout.

La myriadisation : miroir grossissant sur l’annotation manuelle de corpus

Sommaire

1.1	Myriadisation et <i>Crowdsourcing</i> : du jeu dans les termes	5
1.2	Un continuum de pratiques	6
1.3	Des mythes à déconstruire	10
1.4	Myriadisation et sciences participatives	13
1.5	Annotation manuelle de corpus et myriadisation	17

1.1 Myriadisation et *Crowdsourcing* : du jeu dans les termes

J’ai défini le terme *crowdsourcing* et sa traduction « myriadisation » dans de nombreuses publications et cours, mais il me semble indispensable à la lecture et à la compréhension de ce tapuscrit de les clarifier ici.

« Myriadisation » est la traduction de *crowdsourcing* proposée par Gilles Adda (LISN) dans notre article de TALN 2011 (Sagot et al., 2011). Le terme a plu et il est aujourd’hui utilisé par certains collègues, qui le préfèrent aux traductions officielles : « externalisation ouverte »¹ (Office québécois de la langue française) ou « production participative »² (Journal officiel). Si « myriadisation » est moins immédiatement compréhensible que cette dernière proposition, le terme « sonne » aussi bien que son original anglais et projette une idée de masse et de découpage, deux dimensions importantes du *crowdsourcing*. Pour toutes ces raisons et en hommage à Gilles Adda, grâce à qui j’ai fait mes premiers pas dans le domaine de l’éthique en 2010, je privilégie son emploi.

1. Voir : http://gdt.oqlf.gouv.qc.ca/fiche0qlf.aspx?Id_Fiche=45436, où, surprise, « myriadisation » est également proposée.

2. Voir : <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000029331922>.

Chapitre 1. La myriadisation : miroir grossissant sur l’annotation manuelle de corpus

Le mot anglais d’origine, *crowdsourcing*, est un mot valise composé de *crowd* (la foule) et de *outsourcing* (externalisation). Il aurait été créé par un journaliste du magazine Wired, Jeff Howe, en 2006, qui en donne la définition suivante³ :

« Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call. »

Je reviens plus précisément sur ce terme dans le livre que j’ai publié sur le sujet (Fort, 2016), en particulier en ce qui concerne *crowd*, qui ne recouvre pas exactement le sens de « foule » en français. En effet, la définition de *crowd* selon le dictionnaire Merriam Webster⁴ comprend un sens original :

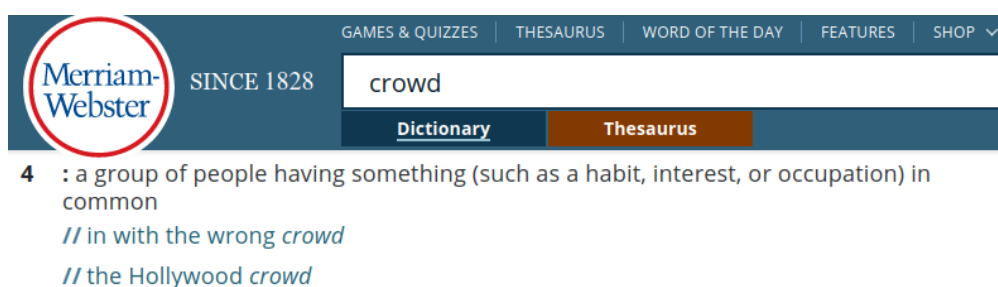


FIGURE 1.1 – Sens 4 de la définition de *crowd* dans le Merriam Webster.

Ce sens du mot, qui met l’accent sur une activité ou un intérêt commun (et qui est relativement négatif, comme en attestent les exemples), n’existe pas dans la définition de la traduction française « foule », selon le TLFi⁵, qui, elle, met l’accent sur la masse.

1.2 Un continuum de pratiques

1.2.1 (Encore) une typologie des myriadisations

Il existe de nombreuses typologies de la myriadisation (voir, notamment Geiger et al. (2011)), souvent rigides dans leur catégorisation. J’ai choisi de montrer le continuum entre les différents types de myriadisation – les sciences participatives, le travail parcellisé et les jeux ayant un but – en les illustrant sur un graphique à

3. L’article d’origine (<https://www.wired.com/2006/06/crowds/>) ne comprend pas de définition, mais Jeff Howe l’a ajouté par la suite sur son blog : <https://crowdsourcing.typepad.com/>.

4. Voir : <https://www.merriam-webster.com/dictionary/crowd>.

5. Voir : <https://www.cnrtl.fr/lexicographie/foule>.

1.2 Un continuum de pratiques

deux dimensions (voir Figure 1.2). Celui-ci permet de rendre compte du positionnement ambigu de certaines plateformes, dont *Phrase Detectives*, qui offre à ses meilleurs participants et par loterie, des bons d'achat sur Amazon (*Chamberlain et al.*, 2009b).

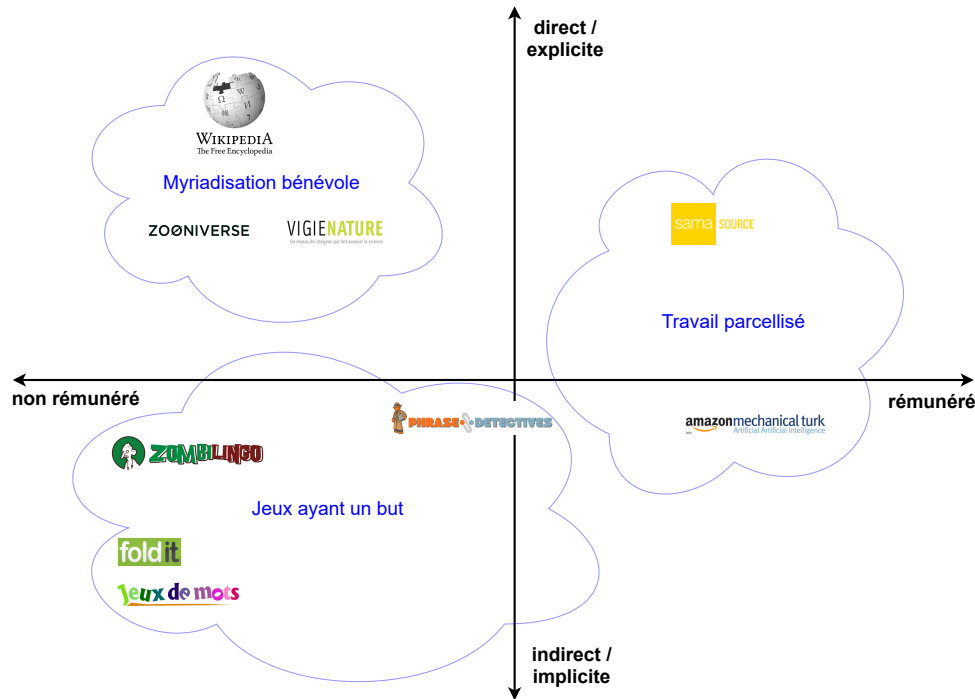


FIGURE 1.2 – Une typologie de la myriadisation selon deux axes : la rémunération et la transparence de la tâche (le participant est-il conscient de ce qu’il construit ?).

Le choix de ces deux dimensions est bien entendu lié à ma réflexion sur l’éthique, puisque le fait de rémunérer ou non une tâche ou de savoir ou non ce à quoi on participe a un impact important sur l’éthique de la plateforme (voir Chapitre 3).

1.2.2 Le travail parcellisé

Le travail parcellisé (*microworking* en anglais) est devenu tellement courant en TAL depuis les années 2010 que le terme *crowdsourcing* est très souvent employé à tort pour le désigner. Ce type de myriadisation consiste à proposer des tâches à des travailleurs en ligne pour qu’ils les réalisent moyennant paiement, *via* une plateforme. La plateforme la plus utilisée pour ce faire est *Amazon Mechanical Turk*, créée par Amazon pour ses propres besoins et ouverte aux demandeurs (*requesters*) en 2005.

Chapitre 1. La myriadisation : miroir grossissant sur l’annotation manuelle de corpus

Le principe est le suivant : les demandeurs proposent des microtâches, ou « Human Intelligence Tasks » (HIT), à des travailleurs en ligne qui les réalisent en échange d’un micropaiement, Amazon fournit l’interface et prélève au demandeur 20 %⁶ du montant versé aux travailleurs. Ce montant peut varier si le demandeur souhaite sélectionner des travailleurs détenteurs de la qualification de *Maîtres* (*Masters Qualification*) ou d’une qualification *Premium* ou s’il souhaite assigner une tâche à plus de 10 travailleurs en parallèle.

Je parle de micropaiement, mais je pourrais sans doute le qualifier de nanopaiement, puisque 90 % des HIT étaient payés moins de 0,1 \$ en 2010 (Ipeirotis, 2010) et que la rémunération plancher acceptée par la plateforme est de 0,01 \$. Des études récentes montrent que les travailleurs perçoivent en moyenne entre 3 \$ de l’heure aux États-Unis et 1,41 \$ de l’heure en Inde (Hara et al., 2019), pour un salaire horaire médian observé de 2 \$ (Hara et al., 2018). On est donc très loin du salaire fédéral minimum américain (pourtant reconnu comme très faible), de 7,25 \$ de l’heure ou de celui de la Californie, à 14 \$ de l’heure. Au-delà des demandeurs abusifs, le travail à la tâche rend difficile de payer correctement les travailleurs. Nous avons ainsi montré que ceux-ci ont tendance à prendre plus de temps pour réaliser une même tâche lorsque celle-ci est mieux payée (Cohen et al., 2016).

Ces tâches sont des microtâches (*microtasks*) qui consistent le plus souvent en des découpages voire des simplifications de tâches réelles, afin de les rendre réalisables par les travailleurs sans formation. En effet, la plateforme ne prévoit pas la possibilité de former les travailleurs, tout au plus de tester certaines de leurs capacités, par exemple leur compréhension de la langue. Un exemple de ce type de simplification est la réduction d’une « vraie » tâche d’annotation en inférences textuelles (inférence, neutre, contradiction) à deux phrases et une question : « Would most people say that if the first sentence is true, then the second sentence must be true ? »⁷ (Bowman et al., 2015).

Bien que cette recherche n’y apparaisse pas, j’ai commencé à travailler pendant ma thèse sur les problèmes que pose Amazon Mechanical Turk en TAL, non seulement du point de vue éthique, mais également de celui de la qualité produite (Fort et al., 2011; Sagot et al., 2011; Fort et al., 2014a). Ces travaux ont été menés en collaboration avec différents collègues, en particulier Gilles Adda et Kevin Bretonnel Cohen, avec qui j’ai par la suite dirigé un numéro spécial de la revue TAL sur l’éthique.

1.2.3 La myriadisation bénévole

Wikipédia est l’exemple type de myriadisation bénévole : les participants

6. Le montant était de 10 % avant 2015.

7. En français : « Est-ce que la plupart des gens dirait que si la première phrase est vraie, alors la deuxième l’est aussi ? ».

créent une encyclopédie en ligne et ils ne sont pas rémunérés pour cela. Ils le font en toute conscience. L'accès à leur production est par ailleurs immédiat, libre et gratuit. Si les données de Wikipédia sont très régulièrement utilisées en TAL, elles n'ont pas été créées dans ce but et la plateforme ne peut donc raisonnablement pas être considérée comme une plateforme de myriadisation pour le TAL.

Lorsque j'ai commencé à m'intéresser à la myriadisation, en 2011, il n'existait pas de plateforme de myriadisation bénévole en ligne pour le TAL. Galaxy Zoo (Lintott et al., 2010) avait démarré en 2007 et était spécialisé dans la classification de galaxies. Le succès du site avait donné naissance à un portail de sciences participatives, Zooniverse, en 2009, mais il ne comprenait et ne comprend toujours que peu de projets liés aux langues (mis à part ceux concernant la transcription de textes). C'est la raison pour laquelle mes collègues du Linguistic Data Consortium (LDC) ont créé récemment LanguageArc (Fiumara et al., 2020), une plateforme de myriadisation bénévole spécialisée dans les langues. LanguageArc présente aujourd'hui une dizaine de projets auxquels les internautes peuvent participer bénévolement et devrait permettre à terme à des chercheurs de créer leur propre projet sur la plateforme, à l'instar de ce qui est proposé sur Zooniverse. Mes collègues, Aurélie Névél (LISN) et Yoann Dupont (ATER à Sorbonne Université), et moi-même avons pu profiter en avance de phase de cette possibilité dans le cadre d'un mémoire de Master 1 (celui de Julien Bezançon) pour faire annoter et produire des stéréotypes en français⁸.

Le problème principal de ce type de plateforme est de parvenir à motiver les participants à venir (motivation), puis à revenir (volition), travailler bénévolement. Les expériences que nous avons menées avec ma doctorante Alice Millour montrent qu'une ludification même légère aide à faire revenir les participants. J'y reviendrai dans le Chapitre 2.

1.2.4 Les jeux ayant un but

Le terme « jeux ayant un but » est la traduction littérale (et peu réussie) de *Games with a purpose*, terme inventé par Luis von Ahn (Von Ahn, 2006), le créateur de ESP Game (Von Ahn and Dabbish, 2004),

Tous les jeux ont un but commun, le divertissement, mais les jeux ayant un but en ont un supplémentaire : créer des données. Ils sont de ce fait parfois considérés comme des jeux sérieux, car ils sont eux-aussi utilisés pour créer autre chose que du divertissement. Cependant, les jeux sérieux ont un but très différent, l'apprentissage. On pourrait même arguer que puisque l'apprentissage génère des données, ce sont les jeux sérieux qui sont une forme de jeux ayant un but.

Au-delà des querelles de typologie, il est tout à fait envisageable de créer des

8. Le projet est accessible après inscription ici : <https://languagearc.com/projects/19>.

jeux qui permettent à la fois d’apprendre et de produire des données. C’est en particulier le but de la plateforme **duolingo**⁹ (Von Ahn, 2013), qui vise à permettre à ses utilisateurs d’apprendre une langue tout en produisant des ressources (en l’occurrence, des traductions). L’action COST *European Network for Combining Language Learning with Crowdsourcing Techniques* (enetCollect), qui comprenait plus de 150 membres de 36 pays, portait également sur ce sujet. J’en ai été porteuse pour la France et responsable du groupe de travail 5 (*Application-oriented specifications for an ethical, legal and profitable solution*) pendant quatre ans (2016–2020).

Je reviens en détail sur les jeux ayant un but dans le Chapitre 2, je ne m’étends donc pas plus ici.

1.3 Des mythes à déconstruire

Quel que soit le type de myriadisation considéré, la vision que la plupart des gens en ont, y compris les chercheurs, tient plus souvent du mythe que de la réalité. Je vais ici tenter de déconstruire les trois principales idées fausses associées à la myriadisation.

1.3.1 Mythe #1 : la myriadisation est un phénomène récent

La plupart des articles traitant de myriadisation pour le TAL citent comme origine des articles tels que Snow et al. (2008) (sur l’utilisation d’*Amazon Mechanical Turk*) pour le travail parcellisé, Von Ahn (2006) pour les jeux ayant un but (du créateur de *ESP Game*) et Lintott et al. (2010) (*Galaxy Zoo*) pour les sciences participatives. Si ces références sont parfaitement adaptées au but recherché (situer la recherche dans le TAL), elles ne reflètent pas la réalité de la myriadisation.

La myriadisation a certes largement profité des possibilités offertes par le Web 2.0 (ou Web social), qui a permis aux internautes, à partir de 2004, d’interagir avec les pages qu’ils visitaient et de ne plus simplement se promener dans un gigantesque musée. De ce fait, la plupart des gens n’imaginent pas la myriadisation sans Internet, il s’agit pour eux d’un phénomène récent, d’une « innovation ».

Or, il est facile de trouver des exemples de sciences participatives bien antérieurs au Web 2.0, à notre siècle, même. Ainsi, le **Longitude prize** a été créé en 1714 par le gouvernement britannique, qui offrait une récompense de vingt mille livres à qui trouverait une méthode simple pour déterminer précisément la longitude d’un navire. La somme, colossale pour l’époque, est revenue à John Harrison, un horloger et charpentier, qui inventa le chronomètre de marine. Il s’agissait bien d’un appel ouvert et on peut tout à fait qualifier cela de myriadisation, même si

9. Voir : <https://fr.duolingo.com/>.

la tâche n'a *a priori* pas été réalisée par une foule de personnes (nous reviendrons sur ce point ci-après).

Un autre exemple de dynamique de sciences participatives, français cette fois, est reflété par l'ouvrage **Instructions pour les voyageurs et les employés des colonies**, dont la première édition a été publiée en 1824 par le Museum National d'Histoire Naturelle (MNHN) pour que les voyageurs et les employés des colonies « [fassent] connaître les résultats de leurs propres expériences, afin d'en profiter et d'en faire profiter le monde savant ». Cet ouvrage répondait visiblement à un besoin et devait permettre d'améliorer les conditions de collecte et de conservation des spécimens récoltés par les expatriés.

Enfin, depuis plus d'un siècle la **Ligue de Protection des Oiseaux (LPO)**¹⁰ assure le suivi des populations d'oiseaux, avec 5 000 bénévoles actifs. Les données ainsi collectées sont utilisées par les chercheurs¹¹ et de nombreuses collaborations ont vu le jour entre la LPO et les équipes de recherche¹².

Les exemples de ce type sont nombreux dans l'histoire des sciences. La frontière entre chercheurs et non-chercheurs (en tant que professionnels) était sans doute plus poreuse qu'aujourd'hui.

1.3.2 Mythe #2 : la myriadisation implique une foule de participants

La myriadisation semble impliquer, de par sa définition même, une foule de participants (*crowd*). Cependant, comme je l'ai montré dans la section 1.1, le mot anglais *crowd* peut référer à un groupe de personnes ayant quelque chose en commun, ce groupe n'étant pas d'une taille nécessairement grande. De fait, nous avons pu montrer dans **Chamberlain et al. (2013)** (pour les jeux) et **Fort (2016)** (plus généralement) que la participation à la myriadisation suit une loi de puissance, typique des activités humaines : peu de personnes participent beaucoup.

L'exemple du jeu ayant un but **Phrase Detectives**¹³ est de ce point de vue prototypique : en un an (de février 2011 à février 2012), treize participants ont gagné la grande majorité des points, donc sont à l'origine de la plupart des annotations produites (voir Figure 1.3). Mieux, un seul de ces joueurs, ancien membre du projet, Livio Robaldo, a tenu la première place du classement pendant plusieurs années et est sans doute à l'origine d'une très grande partie des annotations produites.

10. Voir : <https://www.lpo.fr/partager-vos-observations/partagez-vos-observations>.

11. Voir notamment : http://observatoire-rapaces.lpo.fr/index.php?m_id=20079.

12. Voir par exemple : https://www.patrinat.fr/sites/patrinat/files/atoms/files/2018/12/dutienhat_et_al_2018.pdf.

13. Voir : <https://anawiki.essex.ac.uk/phrasedetectives/>.

Il a été détrôné depuis par Wellington, joueur phare de ZombiLingo ([Guillaume et al., 2016](#)) passé à *Phrase Detectives*.

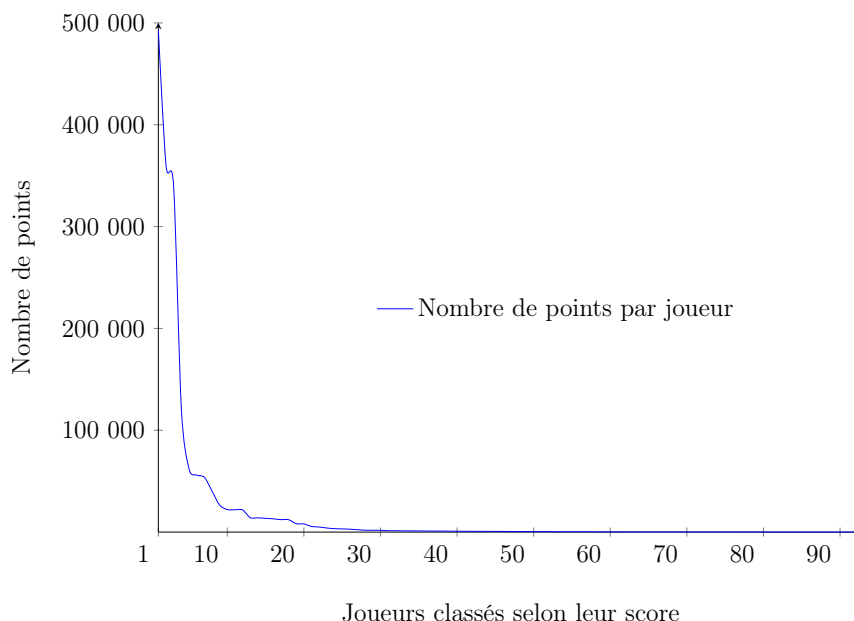


FIGURE 1.3 – Nombre de joueurs sur *Phrase Detectives* selon le nombre de points gagnés dans le jeu (fév. 2011 - fév. 2012) ([Chamberlain et al., 2013](#)).

La notion de foule, on le voit, est donc toute relative. Il en va de même avec la non-expertise des participants.

1.3.3 Mythe #3 : la myriadisation implique des non-experts

Outre que certains très gros joueurs (appelés *whales*¹⁴, dans le monde du jeu) sont des collègues, comme Livio Robaldo, ou même Wellington (dont je garderai l’identité secrète, mais qui est en post-doc en linguistique), les joueurs peuvent devenir de bons experts de la tâche qui leur est proposée. Certains font d’ailleurs preuve d’une compétence linguistique étonnamment fine, comme on peut le constater dans les deux interventions sur le forum de ZombiLingo présentées en Figures 1.4 et 1.5, dans lesquelles les joueurs proposent des tests linguistiques pour justifier leurs critiques. Par ailleurs, dans ce même jeu, certaines personnes se sont spécialisées sur certains phénomènes, comme Dauphine, qui n’a annoté pratiquement que des sujets.

14. Soit « baleines » en français.

1.4 Myriadisation et sciences participatives

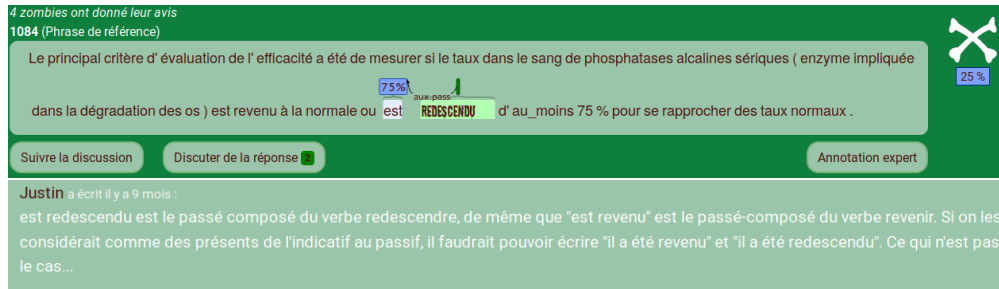


FIGURE 1.4 – Commentaire du joueur Justin dans ZombiLingo à propos de la différence entre le passif et le passé composé dans l’annotation en question.

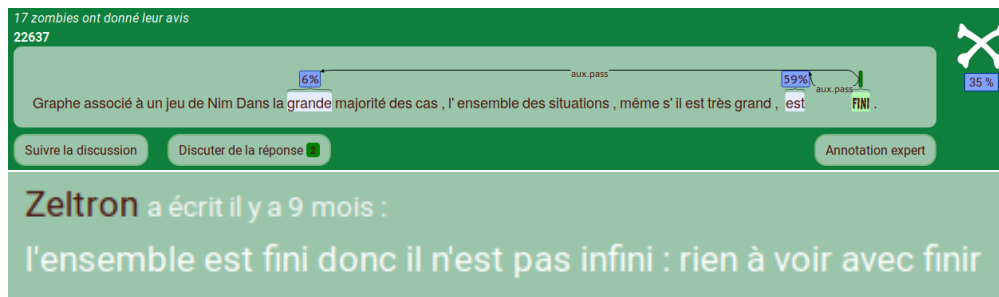


FIGURE 1.5 – Commentaire du joueur Zeltron dans ZombiLingo à propos de « fini », qui n’est pas un verbe au passif dans l’annotation en question.

La myriadisation ne consiste donc pas toujours à profiter d’une foule de non-experts, mais souvent à trouver, et éventuellement former, quelques experts (de la tâche, voire d’une sous-tâche de la tâche) dans la foule.

1.4 Myriadisation et sciences participatives

À partir de 2015, j’ai participé à un certain nombre de groupes de travail ou de projets réunissant des chercheurs travaillant dans le domaine des sciences participatives (voir Figure 1.6). Ces collectifs m’ont permis de réfléchir sous un angle nouveau pour moi à la myriadisation, d’échanger avec des collègues d’horizons très variés, depuis les FabLabs jusqu’au Muséum National d’Histoire Naturelle (MNHN) et de faire évoluer certaines de mes pratiques. J’ai par exemple par la suite proposé des moyens pour recueillir la parole des participants et leur permettre

Chapitre 1. La myriadisation : miroir grossissant sur l’annotation manuelle de corpus

d’interagir entre eux (forum de ZombiLingo).

2015	2016	2017	2018	2019	2020	2021
GPRO SP ATHENA						
		Portail Sciences ensemble				
			Particip-Arc			
					Atelier Sciences ens.	

FIGURE 1.6 – Participation à des groupes de travail sur les sciences participatives (en bleu, CNRS, en vert, Sorbonne Université, en violet, Ministère de la culture).

1.4.1 Pour une définition éthique des sciences participatives

Une large partie du travail mené par ces différents groupes, en particulier le groupe de travail prospectif sur les sciences participatives de l’Alliance ATHENA (GPRO SP) et Particip-Arc, a consisté à définir les sciences participatives et à tenter d’en établir une typologie. Ces efforts n’ont que partiellement abouti, sans doute du fait de la complexité du sujet, mais également de l’hétérogénéité disciplinaire des groupes¹⁵, dont le nombre et la variété des co-auteurs des rapports associés sont témoins¹⁶ et du fait que peu d’entre nous étaient des praticiens.

Ainsi, le GPRO, sous la direction de Sandra Laugier (Professeure de philosophie à Panthéon Sorbonne), a produit un *position paper* (Badouard et al., 2016) qui résulte d’un compromis entre deux visions très différentes des sciences participatives : une vision très large, qui couvre le logiciel libre et Wikipédia, et une vision plus restrictive, qui n’inclut que les travaux impliquant de la recherche. J’étais partisane, avec Romain Julliard, Professeur d’écologie au MNHN, d’une approche plus stricte. Nous avons travaillé ensemble à cette époque là à une définition des sciences participatives que je publie ici pour la première fois, avec son accord :

« Nous définissons [...] les sciences participatives comme des dispositifs à l’initiative de la recherche académique s’appuyant sur la participation de citoyens non chercheurs, volontaires et en conscience, dans un projet de recherche explicite.

15. La multidisciplinarité est certes une richesse, mais elle présente un coût d’entrée très important.

16. Dix-sept co-auteurs pour Badouard et al. (2016), dont des sociologues, des écologues, des informaticiens, des philosophes, des juristes, ... et vingt-sept pour Bernard et al. (2019).

[...] Les sciences participatives se situent au carrefour de trois caractéristiques qui les définissent : elles combinent (i) la recherche, (ii) l'éducation (la formation) et (iii) l'encapacitation (*empowerment*). Ainsi, les jeux ayant un but (*Games With A Purpose*), qui permettent la création de données pour la recherche, *via* des jeux et ce faisant permettent aux joueurs de se former, en font partie, alors que les jeux sérieux, qui ne visent qu'à l'éducation des participants, en sont exclus. De même, les plateformes de création de bien commun telles que Wikipédia n'appartiennent pas aux sciences participatives, puisqu'elles n'ont pas pour objectif de contribuer à la recherche par la participation. [...] Cela dit, ces trois caractéristiques se construisent par la pratique et, dans la galaxie des dispositifs qui se revendiquent ou s'apparentent aux sciences participatives, elles sont inégalement réparties : enrichissement collaboratif de documents scientifiques numérisés, fablab, recherche action participative, savoirs locaux. . .

[...] Les motivations des participants à ces démarches sont variées, mais la co-construction avec la recherche académique d'un bien commun en est une constante. Il s'agit pour les participants d'une activité non lucrative, rémunérée, en quelque sorte, par le divertissement qu'elle offre, le sentiment d'appartenance à une communauté qui partagent les mêmes objectifs. L'éthique est donc au cœur du processus d'appropriation par les participants. Le fait qu'à la différence du travail parcellisé à la **Amazon Mechanical Turk** les participants ne sont pas rémunérés permet de développer le sens de l'intérêt général. Du côté recherche, ce qui est produit par la foule doit revenir à la foule (au moins à la communauté de chercheurs dans son entier), sans quoi l'idée de bien commun perd son sens. »

Cette vision des sciences participatives exclut, volontairement, le travail parcellisé. S'il y a une intersection importante entre myriadisation et sciences participatives, toutes les formes de myriadisation ne peuvent donc pas être considérées comme des sciences participatives et toutes les formes de sciences participatives ne relèvent pas de la myriadisation (par exemple, les FabLab).

Le rapport Houllier ([Houllier and Merilhou-Goudard, 2016](#)), publié un peu avant le *position paper* du GPRO et qui fait maintenant autorité en la matière, reconnaît la variété et la complexité du domaine et propose une vision proche de la nôtre (orientée recherche, donc), mais ne prend parti, ni pour l'ouverture des données, ni contre l'inclusion du microtravail (ou travail parcellisé).

Le projet Particip-Arc ([Bernard et al., 2019](#)) du Ministère de la Culture incluait des participants différents du GPRO (à l'exception de quelques uns, dont le MNHN et moi-même), nous avons donc dû repasser par une phase définitoire et

Chapitre 1. La myriadisation : miroir grossissant sur l’annotation manuelle de corpus

typologique. Ces travaux, très riches, n’ont malheureusement pas donné lieu à une publication, tant la diversité des sciences participatives mériterait une thèse pour la représenter et l’organiser de manière satisfaisante. Cependant, le projet a fait l’objet d’un numéro spécial de la revue Culture et Recherche, qui présente de belle manière la richesse de nos travaux¹⁷.

1.4.2 Les données pas données des sciences participatives

Dans ce cadre, j’ai travaillé avec Lisa Chupin (Université Paris Descartes / DICEN-IDF) sur les données des sciences participatives (Chupin and Fort, 2019), ce qui m’a permis d’approfondir la question de la caractérisation et de l’ouverture de celles-ci.

Nous nous sommes intéressées aux données des projets présentés dans le cadre de Particip-Arc, depuis les observatoires du MNHN¹⁸, les Herbonautes¹⁹ (transcription d’étiquettes d’herbiers), le projet Testaments de poilus²⁰ (transcription), jusqu’à nos projets ZombiLingo et Recettes de grammaire, ainsi qu’à celles des plateformes Zooniverse²¹ et Tela Botanica²² (réseau des botanistes francophones).

Des données imprévues

Il en ressort qu’outre les données prévues au départ, ces plateformes collectent deux autres types de données. Le premier correspond aux données de connexion des participants : identifiant, mot de passe, profil éventuel. Ces données peuvent être identifiantes et donc poser des problèmes de droit et d’éthique. Viennent ensuite les données que nous avons qualifiées d’« incidentes », comme les commentaires dans les forums ou les mails de participants. Ces dernières ont souvent un statut légal et scientifique incertain, en particulier lorsque la collecte n’a pas été planifiée (ce qui est souvent le cas). Or, elles peuvent se révéler très intéressantes, notamment pour expliquer certaines décisions, voire corriger les données de référence. Nous avons ainsi pu corriger des erreurs dans huit phrases de référence de ZombiLingo grâce aux retours des joueurs dans le forum (Fort and Guillaume, 2019).

Enfin, les contributions des participants forment des données intermédiaires qu’il faudra dans la plupart des cas faire valider par d’autres participants, soit directement, soit indirectement par une ou plusieurs contributions concurrentes (il

17. Voir : <https://lstu.fr/revue-culture-et-recherche>.

18. Voir : <https://www.vigienature.fr/>.

19. Voir : <http://lesherbonautes.mnhn.fr/>.

20. Voir : <https://testaments-de-poilus.huma-num.fr/>.

21. Voir : <https://www.zooniverse.org>.

22. Voir : <https://www.tela-botanica.org/>.

faudra alors sélectionner la meilleure contribution selon un algorithme quelconque) pour obtenir les données finales.

Des données plus ou moins ouvertes

La majorité des sites analysés (Zooniverse, certains sites du MNHN) ne propose pas de télécharger directement les données récoltées. L'accès en est limité et fonction de la volonté des équipes de recherche, ce qui n'empêche pas leur utilisation, mais la rend moins immédiate.

Certains sites ne proposent que la consultation sur la plateforme et ne rendent pas les données elles-mêmes disponibles.

À l'inverse, Tela Botanica permet aux participants de décider eux-mêmes de mettre ou non à disposition leurs données (y compris leur profil).

J'ai choisi, avec mes collègues et ma doctorante, de mettre les données produites après validation ou agglomération, à disposition de tous, directement sur le site de myriadisation, selon le principe de « ce qui est produit par la foule doit revenir à la foule »²³. Nous ne divulguons par contre aucune autre donnée, en particulier nous ne publions pas les données de connexions ou les profils utilisateurs, que nous réduisons au minimum de toutes façons, selon le principe éthique de minimisation²⁴.

1.5 Annotation manuelle de corpus et myriadisation

1.5.1 Annoter, c'est quantifier

Benoît Habert (aujourd'hui retraité, à l'époque ENS Lyon), dans le rapport qu'il a rédigé sur ma thèse, citait abondamment Alain Desrosières, un historien des statistiques. Il m'a fallu du temps pour saisir tout l'intérêt des travaux de celui-ci et leur lien avec le TAL.

Alain Desrosières a travaillé sur les catégories, en particulier sur les fameuses catégories socio-professionnelles (CSP) de l'INSEE. Il a dans ce cadre établi certaines distinctions utiles au TAL. Il fait notamment la distinction entre la mesure et la quantification. En effet, certaines réalités sont immédiatement mesurables, comme la hauteur du Mont Everest (8 848 m), d'autres ne le sont pas, par exemple le nombre de chômeurs, car il faut auparavant se mettre d'accord sur *ce qu'est un chômeur* :

23. Cette maxime me vient de Mathieu Lafourcade (LIRMM), créateur de Jeux de Mots : <http://www.jeuxdemots.org/>.

24. Voir : <https://www.cnil.fr/fr/definition/minimisation>.

Chapitre 1. La myriadisation : miroir grossissant sur l’annotation manuelle de corpus

« Mais précisément la définition et la mesure de la population active et du chômage relèvent d’une autre épistémologie que celle de l’étoile polaire. Elles impliquent des conventions (analogues aux principes généraux des lois et des codes votés par les Parlements) et des décisions (analogues à celles d’un juge) d’affecter tel cas à telle classe. » (Desrosières, 2001)

Il décrit par ailleurs ce processus de classement, le *codage* et pointe du doigt le cœur du problème, l’équivalence supposée entre les éléments d’une même classe (ou catégorie) :

« Un codage est une décision conventionnelle de construire une classe d’équivalence entre divers objets, la ‘classe’ étant jugée plus ‘générale’ que tout objet singulier. La première condition pour cela est de supposer que tous ces objets peuvent être comparés, ce qui ne va pas de soi » (Desrosières, 1989)

Pour manipuler ces réalités, qui ne sont pas de l’ordre de la mesure mais de la quantification, il faut donc d’abord établir des conventions :

« Quantifier, c’est se mettre d’accord, puis mesurer » (Desrosières, 2008)

Le lien avec l’annotation est évident : l’annotation est un processus visant à ajouter des informations *interprétatives* (Leech, 1997; Habert, 2005) sur du texte, souvent des catégories, qu’il faut décrire dans un guide d’annotation et dont il faut évaluer la compréhension et l’application par les annotateurs par le biais de mesures d’accord inter-annotateurs (Artstein and Poesio, 2008).

Annoter, c’est donc quantifier. Annoter, c’est se mettre d’accord. Or, si cela est loin d’être simple dans un cadre traditionnel (ma thèse s’en fait l’écho), ça l’est encore beaucoup moins dans un contexte myriadisé, où :

- on ne connaît *a priori* pas les annotateurs,
- ils ne sont *a priori* pas spécialistes du domaine,
- la communication est asynchrone (voire inexistante).

La myriadisation a donc un effet loupe sur les difficultés de l’annotation, qu’elle nous oblige à affronter. De ce fait, la manière dont nous avons posé les problèmes et les solutions que nous avons développées pour myriadiser pourraient sans aucun doute bénéficier à l’annotation traditionnelle. J’y reviens ici de manière analytique et dans le Chapitre 2 pour la partie expérimentale.

1.5.2 Une activité (insuffisamment) outillée

L’activité d’annotation, même traditionnelle, est très majoritairement aujourd’hui une activité outillée. Il existe pléthore d’outils d’aide à l’annotation, à tel

point que depuis le recensement que j'en avais fait pendant ma thèse, un outil de recherche a été créé, **Annotationsaurus** (Neves and Seva, 2019), qui en liste 115 sur le site GitHub correspondant²⁵. Cependant, parmi les 78 outils passés au crible dans Neves and Seva (2019), seuls trois proposent le calcul de l'accord inter-annotateurs (**WebAnno**, **Djangology** et **LightTag**), un élément clé de la gestion d'une campagne d'annotation. Cela montre à quel point cette activité reste mal connue.

Si **WebAnno** (De Castilho et al., 2014) est l'outil satisfaisant le plus de critères dans leur étude et s'il est selon moi celui qui offre le meilleur compromis fonctionnalités/courbe d'apprentissage²⁶, il ne propose que deux types d'adaptation de l'interface pour rendre la tâche moins complexe : i) le masquage de certaines couches d'annotation (permettant par exemple de masquer la couche de catégories grammaticales lorsque l'on annote la syntaxe) et ii) la possibilité de pré-annoter le corpus. Aucun des outils évalués dans Neves and Seva (2019) ne va d'ailleurs plus loin à ma connaissance. En particulier, aucun ne propose une ré-organisation de la tâche d'annotation en fonction de ses dimensions de complexité telles que nous les avons identifiées dans Fort et al. (2012) ou sous une autre forme.

Les interfaces ludifiées et les jeux ayant un but que j'ai participé à créer, notamment **ZombiLingo** et **Recettes de grammaires**, ont été conçus en prenant en compte ces dimensions de complexité et ont permis la réalisation des tâches qu'ils visaient par myriadisation volontaire. Ainsi, dans **ZombiLingo** les annotateurs annotent par relation et non par phrase, ce qui leur permet de se former par étape, voire de se spécialiser (voir la figure 1.7).

Cette avancée dans la conception des outils d'annotation pourra je l'espère inspirer les créateurs d'outils pour l'annotation traditionnelle. Une collaboration avec les créateurs de **WebAnno** (et maintenant de son successeur **Inception** (Klie et al., 2018)) pourrait être bénéfique de ce point de vue.

1.5.3 Des annotateurs sous influences

Si les biais liés aux stéréotypes représentent aujourd'hui un sujet de recherche à part entière dans le TAL de l'ère du *Deep learning* (j'en parle dans le Chapitre 3), c'est d'un tout autre type de biais dont il est question ici : les influences auxquelles sont soumises les annotateurs humains dans le cadre de leur tâche d'annotation.

En effet, il a été démontré que les annotateurs sont influencés, d'une part par les outils d'aide à l'annotation utilisés (Dandapat et al., 2009) et d'autre part par les pré-annotations (produites par des systèmes de TAL) qu'ils doivent corriger (Fort and Sagot, 2010). Les mêmes expériences ont montré que les annotateurs

25. Voir : <https://github.com/mariananeves/annotation-tools>.

26. **WebAnno** est l'outil que j'utilise dans mon cours sur l'annotation.



FIGURE 1.7 – Page de sélection des relations à annoter dans ZombiLingo.

bien formés ou plus expérimentés sont moins sensibles aux biais. Or, l’importance de la courbe d’apprentissage des annotateurs est connue depuis longtemps, puisque les créateurs du *Penn Treebank* l’ont évaluée, dans les années 90, à un mois pour l’annotation en parties du discours et à deux mois pour l’annotation en syntaxe (Marcus et al., 1993).

Une autre influence, qui à ma connaissance n’a pas été validée scientifiquement dans le cadre de l’annotation, mais que j’ai pu observer lors des expériences que j’ai menées pendant ma thèse et qui a également été constatée par des collègues du Linguistic Data Consortium (LDC) est celle des annotateurs entre eux : certains annotateurs parviennent à imposer leurs décisions aux autres, qu’ils aient raison ou non.

Dans l’annotation traditionnelle, les annotateurs travaillent donc sous des influences diverses, plus ou moins bien identifiées, avec des conséquences plus ou moins graves sur la qualité de l’annotation produite.

Dans l’annotation myriadisée, l’influence des annotateurs entre eux est en général très peu présente, car peu de plateformes leur proposent des moyens de communiquer. Sur le forum que nous avons ajouté à ZombiLingo, les joueurs ont surtout communiqué avec les gestionnaires de la plateforme (Bruno Guillaume et moi-même) et peu entre eux : nous n’avons identifié que dix cas d’échanges entre joueurs sur 205 discussions (Fort and Guillaume, 2019).

En revanche, le travail d’annotation est en général simplifié, donc outillé (y

compris pré-annoté), ce qui soumet les participants à un biais de confirmation important. Il faut en être conscient et tenter de le contrebalancer, par exemple par des phrases-tests ou des « pièges », conçus pour maintenir l’attention des joueurs (voir Chapitre 2). Cette vérification de l’attention des annotateurs est réalisée de manière moins systématique dans l’annotation traditionnelle, ce qui est à mon sens une erreur. Le calcul de l’accord intra-annotateur (entre une annotation réalisée à un instant t et la même annotation, par le même annotateur, à l’instant $t+10$ *jours*, par exemple), qui participe du même effort, est également trop peu effectué.

1.5.4 Le mythe de l’annotateur expert

Bien qu’elle soit souvent mentionnée dans les articles de recherche en TAL qui incluent de l’annotation manuelle, l’expertise des annotateurs est rarement approfondie ou traitée en tant que telle. On mentionne des « experts », parfois opposés à des annotateurs « naïfs » ou « non-experts », éventuellement en ajoutant des détails sur leur formation (Marcus et al., 1993) ou en spécifiant qu’ils sont des « experts du domaine » (Candito et al., 2014). Si Amber Stubbs fait une distinction entre linguistes et experts, qui sont dans son cas des docteurs en médecine ou des spécialistes en biomédical (Stubbs, 2012), celle-ci est limitée à un domaine spécifique.

Cette question que l’on évacue un peu rapidement est pourtant d’importance, car elle influence toute la campagne d’annotation, depuis la sélection des annotateurs et leur formation jusqu’à l’outillage proposé. Or, elle est loin d’être triviale et pose la question de ce qu’on annote et dans quel but, exactement.

Lorsque, dans le cadre du programme Quæro, nous avons annoté un corpus de presse ancienne en entités nommées (Rosset et al., 2012), nous avons été confrontés à une difficulté inattendue, celle des limites de notre culture, en particulier historique. Ainsi, en présence du terme « krach Macé », nous étions incapables de dire s’il s’agissait du nom d’une personne, d’un événement ou même d’une entreprise. Il en a été de même pour des expressions comme « mandement de Carême » (voir la figure 1.8).

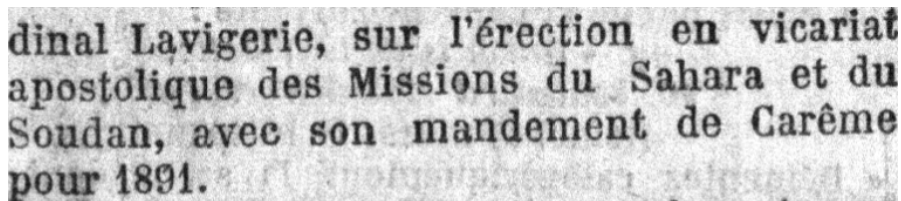


FIGURE 1.8 – Extrait du corpus de presse ancienne de la campagne Quæro, à annoter en entités nommées (Rosset et al., 2012).

Chapitre 1. La myriadisation : miroir grossissant sur l’annotation manuelle de corpus

L’efficacité de l’annotation en a pâti, puisque nous avons dû effectuer des recherches, souvent infructueuses d’ailleurs, sur le Web. Dans ce cas particulier, est-ce que nous étions réellement des experts ? Est-ce qu’un historien spécialiste de la période concernée n’aurait pas été plus efficace que des chercheurs en TAL ? Est-ce qu’il n’aurait pas été plus rapide de le former à notre guide d’annotation plutôt que de faire toutes ces recherches chronophages ?

Prenons un autre exemple, davantage lié à l’actualité, celui d’un corpus de pharmacologie comme EMEA dans le corpus Sequoia ([Candito and Seddah, 2012](#)), annoté en syntaxe, dont voici un extrait :

« Pour les SCA, la durée de la perfusion dépend de la manière dont le SCA doit être traité : elle peut durer jusqu’à 72 heures au maximum chez les patients devant recevoir des médicaments. »

Qui serait un expert pour annoter ce sous-corpus en syntaxe ? Un linguiste (de quel type ? de quel niveau ?) ? Un pharmacien ? Un taliste ? Est-ce qu’un joueur sur un jeu ayant un but pourrait annoter cette phrase de manière satisfaisante ?

Au-delà de ces exemples, c’est à mon avis la question de l’expertise qu’il faut poser : qu’est-ce qu’un expert ? L’expérience de la myriadisation confronte directement à cette question, que j’ai abordée par ce prisme dans [Fort \(2017\)](#). J’ai déjà montré dans la section 1.3.3 que les participants ne sont pas forcément les non-experts que l’on croit, mais bien plutôt des experts de la tâche proposée, qui parviennent à annoter des phrases complexes pourvu qu’elles soient proposées dans un cadre décomplexifié, comme dans ZombiLingo (voir la figure 1.9).

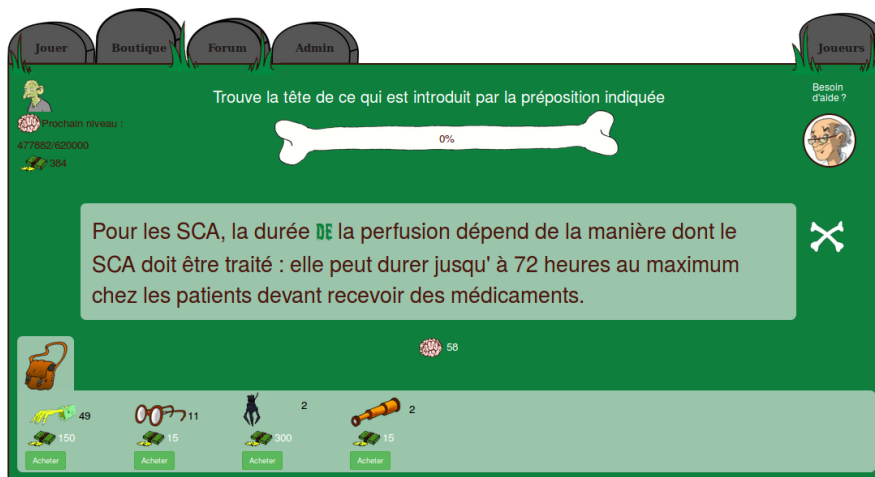


FIGURE 1.9 – Phrase d’EMEA à annoter par étapes dans ZombiLingo : ici, la bonne réponse est « perfusion ».

Nous avons montré dans [Fort et al. \(2017b\)](#) que les joueurs sont capables d’annoter en syntaxe de dépendances un corpus complexe (plus grande complexité lexicale, plus grande profondeur moyenne des dépendances et plus longues relations). Leurs résultats s’étagent entre 0,71 et 0,96 de F-mesure en fonction des relations, avec des scores inférieurs à ceux de l’analyseur syntaxique *Talismane* ([Urieli, 2013](#)) pour les dépendances les plus simples, comme déterminant (DET), mais supérieurs dans les cas de relations complexes comme les coordinations (COORD et DEP.COORD). Ces résultats sont intéressants, en particulier pour améliorer l’apprentissage des outils sur les dépendances complexes. Cependant, il faut noter que le corpus de référence utilisé était de petite taille (39 phrases, 1 245 tokens) et que les outils produisent de meilleurs résultats depuis l’avènement de l’apprentissage profond (voir par exemple UDPipe 2 ([Straka, 2018](#); [Straka et al., 2019](#))). Il serait donc utile de renouveler l’expérience en l’élargissant.

La myriadisation pose donc directement la question de l’expertise des annotateurs et de l’adaptation de l’outillage à leurs capacités, question qui est souvent passée sous silence dans l’annotation traditionnelle, générant perte de temps et de qualité produite.

J’ai défini dans ce chapitre ce qu’est la myriadisation, ce qu’elle n’est pas, et comment elle peut apporter un point de vue enrichissant pour l’annotation traditionnelle en TAL. Je me suis appuyée pour cela sur les réflexions typologiques que j’ai menées dans différents collectifs et sur mon expérience de la myriadisation par le jeu, que je vais détailler dans le chapitre suivant.

Les jeux ayant un but : un modèle encore incertain

Sommaire

2.1	Les différents types de jeux ayant un but	25
2.2	Ludifier la syntaxe en dépendances : ZombiLingo	31
2.3	Des jeux aux plateformes (moins) ludifiées	37
2.4	Les incertitudes du modèle	45

J’ai positionné, dans le chapitre 1 les jeux ayant un but parmi les différents types de myriadisation. Je me propose dans ce chapitre de les définir de manière plus approfondie et de montrer à la fois ce qu’il est possible d’obtenir comme résultats et les défis que présente la création de tels jeux.

Il est à noter que le livre de Mathieu Lafourcade (LIRMM) sur les jeux ayant un but ([Lafourcade et al., 2015a](#)) fournit une liste détaillée de jeux et je suggère de s’y référer pour davantage d’exemples.

2.1 Les différents types de jeux ayant un but

J’ai proposé dans mon livre sur l’annotation collaborative de corpus ([Fort, 2016](#)) une classification des jeux ayant un but en trois types principaux, que je reprends ici.

Les deux premiers types de jeux sont assez classiques : l’un, le plus courant, fait appel aux connaissances du monde et de la langue des locuteurs, l’autre à leurs connaissances scolaires. Le troisième type est moins évident et consiste à faire appel aux capacités d’apprentissage des locuteurs. La grande majorité des jeux ayant un but appartient aux deux premiers types.

2.1.1 Les jeux faisant appel aux connaissances du monde et de la langue des locuteurs

Le premier jeu ayant un but, ESP Game, a semble-t-il¹ été créé par Luis von Ahn en 2004 (Von Ahn and Dabbish, 2004). Il consistait, pour les joueurs, à regarder une image et à entrer des mots correspondant à cette image en parallèle d'un autre joueur. Si les deux joueurs avaient proposé les mêmes mots, ils gagnaient des points (voir la figure 2.1).



FIGURE 2.1 – Figure 1 de von Ahn and Dabbish (2004) : deux participants jouant sur ESP Game et ayant trouvé la même étiquette (« purse »).

Ce jeu ne nécessitait pas de formation particulière et ne demandait aux participants qu'une connaissance de base du monde et de la langue. Son succès fut massif, puisqu'il permit de récolter plus de dix millions d'étiquettes d'images en quelques mois (Von Ahn, 2006).

ESP Game a inspiré Mathieu Lafourcade pour JeuxDeMots² (Lafourcade, 2007), qui reprend l'idée du duel entre joueurs et certains mécanismes du jeu, dont les mots tabou (ceux qui ont déjà été beaucoup joués et sont à éviter). Il consiste à entrer dans un champ texte le plus de termes possibles correspondant à la consigne donnée pour un terme donné et dans un temps restreint (voir la figure 2.2). Les termes sont ensuite comparés à ceux entrés par un autre joueur et s'il y a correspondance, des points sont alloués aux deux. Là encore, le jeu a eu beaucoup de succès et en dix ans il a permis la création d'un réseau lexical du français qui évolue

1. Il est fort possible que des jeux de ce type aient été utilisés avant le Web 2.0, mais nous n'en n'avons pas trace.

2. Voir : <http://www.jeuxdemots.org>.

2.1 Les différents types de jeux ayant un but

constamment, grossissant d'environ 20 000 termes et 1,4 million de relations par mois (Lafourcade et al., 2018).

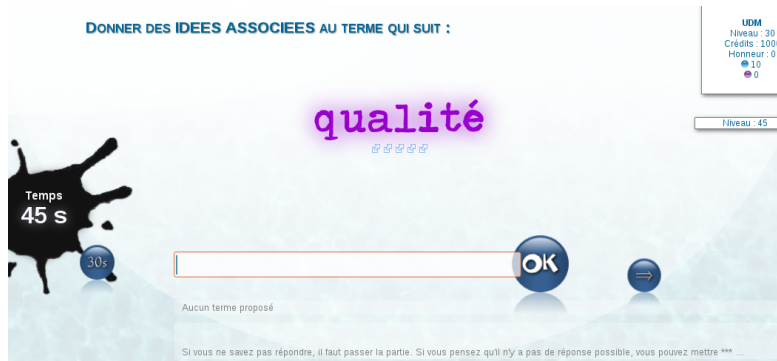


FIGURE 2.2 – Interface de **JeuxDeMots** : le joueur doit trouver des idées (termes) associées à « qualité ».

JeuxDeMots a la particularité d'être l'un des rares vrais *jeux* pour le TAL et non une interface plus ou moins ludifiée. Il a beaucoup évolué avec le temps et propose aux joueurs une véritable immersion, avec de nombreuses fonctionnalités qu'ils découvrent au fur et à mesure (des cadeaux, des patates chaudes, des mots à capturer, etc).

Une autre particularité de ce jeu est qu'il est accompagné d'une galaxie de petits jeux, qui se sont multipliés au fil du temps : **TierXical**, **SexIT** (Lafourcade and Fort, 2014), **PolitIT** (Tisserant and Lafourcade, 2015), **Totaki**, **LikeIT** (Lafourcade et al., 2015b), **Emot** (Lafourcade et al., 2016), **ColorIT** (Lafourcade et al., 2014), **AskYou**, **AskIT**, **Sélemo**, **Top10**, **Yakadiroù**. Ces douze jeux, dont dix de vote, sont disponibles sur un portail accessible depuis le jeu principal³. Ils permettent d'enrichir la ressource créée par **JeuxDeMots**, principalement en ajoutant des relations ou en renforçant certains liens dans le réseau. Leur simplicité n'a d'égal que leur efficacité et j'ai participé à les mettre en valeur dans un article commun publié à LREC en 2018 (Fort et al., 2018b). Mon but était de pousser à une adoption plus large de ce type de jeux, par exemple par le biais d'une plateforme accessible à des linguistes pour y déposer leurs projets (une question, les réponses possibles, un lexique ou des phrases d'entrées, quelques éléments d'illustration), qui seraient transformés automatiquement en jeux de vote. J'ai échoué à faire financer ce projet par le CNRS *via* l'appel Mastodons en 2015 et par la Sorbonne en 2016 et je n'ai pas trouvé d'opportunité depuis.

3. Voir : http://imaginat.name/JDM/Page_Liens_JDMv2.html.

2.1.2 Les jeux faisant appel aux connaissances scolaires des locuteurs

La plupart des jeux présents sur le portail LingoBoingo⁴ du Linguistic Data Consortium (LDC) sont plutôt des interfaces ludifiées et nécessitent un peu plus de connaissances, *a minima* des connaissances scolaires, pour y jouer.

Parmi ceux-ci, **Phrase Detectives**⁵ est le plus ancien (Chamberlain et al., 2008; Poesio et al., 2013), puisqu'il a été mis en ligne sans doute peu après JeuxDeMots. Cette interface ludifiée permet aux participants d'une part d'annoter des liens de co-référence (très majoritairement d'anaphore) dans des textes issus de Wikipédia (voir la figure 2.3) et d'autre part de valider (ou non) les annotations des autres joueurs (voir la figure 2.4).

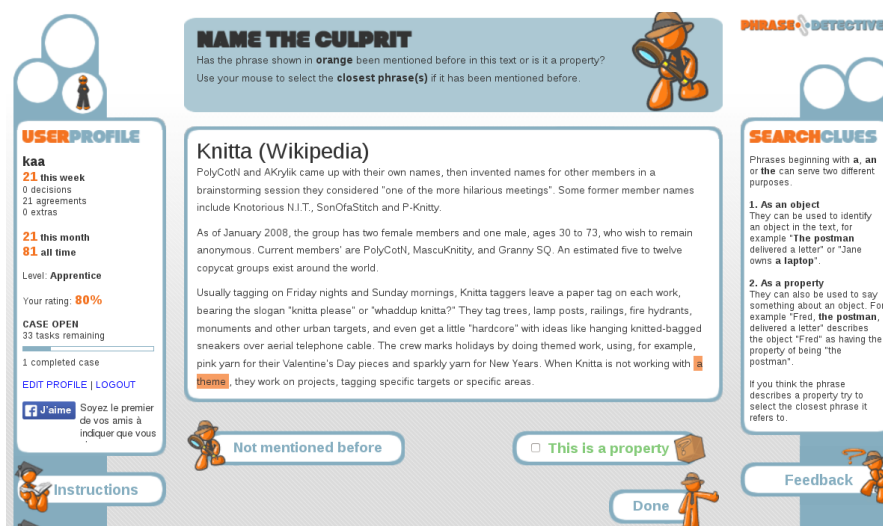


FIGURE 2.3 – Interface de **Phrase Detectives** : le joueur doit trouver un antécédent de « a theme » (s'il existe).

Avant de pouvoir réaliser ce type d'annotation ou de validation, le joueur doit passer par une phase de formation, afin de se familiariser avec la tâche. La résolution d'anaphore est une tâche qu'on apprend à l'école primaire, il s'agit donc d'une connaissance scolaire.

Deux mécanismes sont mis en œuvre pour s'assurer de la qualité de la production (Chamberlain et al., 2008). D'une part, les joueurs se voient proposés aléatoirement des textes de référence à annoter et ils perçoivent un bonus de points s'ils réussissent, ou ont un retour négatif s'ils se trompent (« *comparative scoring* »).

4. Voir : <https://lingoboingo.org/>.

5. Voir : <http://anawiki.essex.ac.uk/phrasedetectives>.

2.1 Les différents types de jeux ayant un but



FIGURE 2.4 – Interface de **Phrase Detectives** : le joueur doit (in)valider l’annotation d’un autre joueur entre « Knitta taggers » et « The group ».

D’autre part, si, lors de la phase de validation, un autre joueur valide leur annotation, ils gagnent des points (« *collaborative scoring* »). Ce dernier mécanisme a pour conséquence que les joueurs peuvent gagner des points sans jouer, simplement parce que d’autres joueurs ont validé leurs annotations. C’est une caractéristique intéressante du jeu, car cela peut pousser à revenir sur la plateforme pour vérifier si l’on a gagné des points entre temps.

Phrase Detectives a ceci de particulier que les participants peuvent gagner des bons d’achat **Amazon**, 50 £ pour le meilleur score du mois, 30 £ pour le meilleur commentaire, par exemple. Il se situe donc dans un continuum entre le travail parcellisé et les jeux ayant un but.

Le jeu est un succès, puisqu’au 5 avril 2022, 61 639 personnes y ont participé, produisant 5 352 830 annotations⁶, d’une qualité tout à fait satisfaisante de plus de 83 % d’accord observé entre un expert et les joueurs ([Chamberlain et al., 2009a](#)).

6. Massimo Poesio, communication personnelle, le 5 avril 2022.

2.1.3 Les jeux faisant appel aux capacités d'apprentissage des locuteurs

Une dernière catégorie de jeux ayant un but regroupe ceux qui font appel à la volonté, au plaisir même, qu'ont les participants d'apprendre des règles et de les respecter. Si l'on se réfère à la catégorisation de Roger Caillois, cela correspond au *ludus*, qui est soumis à des conventions arbitraires et qui pousse à l'acquisition d'une maîtrise pour résoudre une « difficulté créée à dessein » (Caillois, 1958, p. 55)⁷. Ce type est opposé à la *paida*, le jeu libre. En créant un jeu présentant des règles complexes, nous profitons donc de la capacité et du plaisir d'apprendre des participants.

Bien entendu, les autres types de jeux ayant un but ont une dimension *ludus*, puisqu'ils ont également des règles, mais ils ne nécessitent pas un apprentissage, une maîtrise, du même ordre. Encore une fois, il s'agit d'un continuum et la classification que je présente ici doit être entendue comme une progression.

Le premier jeu de ce type est à ma connaissance **FoldIt** (Cooper et al., 2010), un jeu en trois dimensions de repliement de protéines (voir Figure 2.5). Ce jeu permet à des non spécialistes en biochimie d'apprendre à replier des protéines, tâche pour laquelle l'excellente vision en trois dimensions de l'œil humain nous rend particulièrement efficaces.

La formation des participants a lieu par étapes, par le biais d'un tutoriel décomposé par concept, chaque concept comprenant un ensemble de puzzles qu'il faut réussir pour pouvoir accéder aux puzzles suivants.



FIGURE 2.5 – Formation par étapes dans FoldIt (à gauche) et interface du jeu (Figure 4 de (Cooper et al., 2010)).

Ce jeu a eu beaucoup de succès et a permis de résoudre la structure cristalline de la protéine responsable de la propagation du virus du SIDA chez les macaques rhésus (Khatib et al., 2011). Ce problème était non résolu depuis une dizaine

7. Je remercie ici Didier Ozil (doctorant en sociologie à Montpellier P. Valéry) qui m'a fait découvrir cet ouvrage.

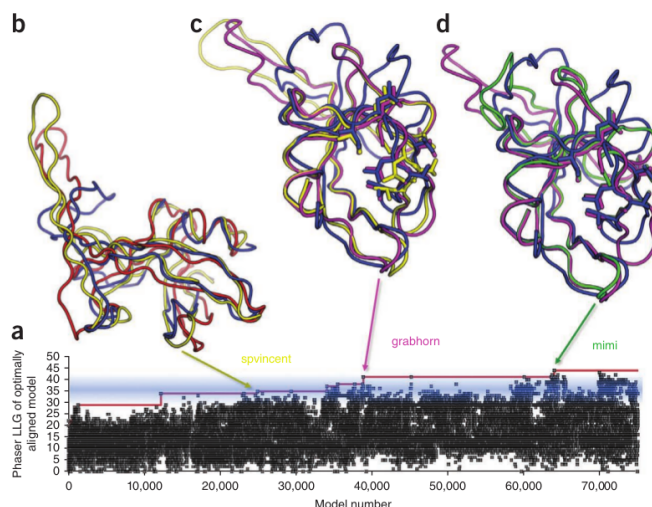


FIGURE 2.6 – Progression des trois joueurs de FoldIt vers la solution (Figure 2 de (Khatib et al., 2011)).

d’années et la solution a été trouvée en quelques semaines par une équipe de trois joueurs (voir Figure 2.6).

Nous nous sommes inspirés de FoldIt, ainsi que de *Phrase Detectives*, pour développer ZombiLingo (Fort et al., 2014b; Guillaume et al., 2016), un jeu d’annotation en syntaxe de dépendances.

2.2 Ludifier la syntaxe en dépendances : ZombiLingo

ZombiLingo est en partie le résultat d’un pari. En effet, l’annotation en syntaxe de dépendances est l’une des tâches d’annotation les plus complexes du TAL et il m’a semblé intéressant de tester la grille de complexité mise au point pendant ma thèse sur cette tâche, afin de la rendre réalisable par le biais d’un jeu ayant un but. Bien entendu, le jeu correspondait également à un besoin, le français manquant encore à l’époque d’un corpus de grande taille librement disponible et redistribuable⁸. J’ai donc commencé à travailler sur ZombiLingo avec Bruno Guillaume à partir de 2014.

8. Le corpus arboré de Paris 7 ou « French Treebank » (Abeillé et al., 2003) n’était pas librement redistribuable et Sequoia (Candito and Seddah, 2012) était de taille réduite.

2.2.1 Une tâche complexe

L’expression « jeu d’annotation en syntaxe de dépendances » fait figure d’oxymore, tant la syntaxe en dépendances n’est pas une activité *a priori* ludique. En effet, non seulement le guide d’annotation du corpus Sequoia (Candito et al., 2014) que nous avons utilisé est long (50 pages) et complexe (29 types de relations), mais les décisions à prendre sont souvent contre-intuitives. Ainsi, à la différence du schéma d’annotation des corpus plus récents du projet Universal Dependencies (UD)⁹, les relations ne s’appuient en général pas sur le mot sémantiquement plein. Dans l’exemple présenté en Figure 2.7, le dépendant de la relation `a_obj` (objet introduit par « à ») est donc la préposition (« au »).

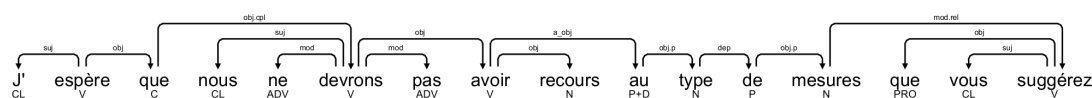


FIGURE 2.7 – Exemple d’annotation en syntaxe de dépendances selon le schéma d’annotation du corpus Sequoia.

2.2.2 Décomplexifier n’est pas simplifier

Afin de rendre cette tâche jouable, nous l’avons décomplexifiée. Pour cela, nous avons identifié les points les plus complexes de l’annotation grâce à la grille d’analyse de complexité développée pendant ma thèse, avec Adeline Nazarenko (LIPN) et Sophie Rosset (LISN) (Fort et al., 2012). Ces points « durs » concernaient principalement l’expressivité du langage d’annotation (langage relationnel d’arité 2), la taille du schéma d’annotation et la discrimination (identification des segments à annoter).

Nous avons ensuite cherché des solutions pour contourner les problèmes ou les alléger. Nous avons ainsi décidé de proposer de jouer (d’annoter) par type de relation (par exemple, `sujet`), plutôt que par empan de texte (la phrase), comme le montre la figure 2.8. Cela limite la complexité due à la taille du schéma d’annotation en permettant aux joueurs de se concentrer sur un phénomène à la fois. Certains joueurs se spécialisent d’ailleurs sur une ou plusieurs relations et ne jouent pas les autres. C’est le cas par exemple de Dauphine, qui a surtout joué des relations `sujet`.

Nous avons également choisi de pré-annoter le texte et de proposer aux joueurs d’identifier uniquement la tête ou le dépendant d’une relation (voir Figure 2.9) plutôt que la relation entière, afin de réduire la complexité due à l’expressivité du

9. Voir : <https://universaldependencies.org/>.

2.2 Ludifier la syntaxe en dépendances : ZombiLingo



FIGURE 2.8 – Interface de sélection de relations à jouer dans ZombiLingo.

langage d’annotation et à la discrimination. Au cas où la pré-annotation proposée (le couple relation / élément surligné) est fausse, le joueur doit cliquer sur la croix d’os à droite de la phrase (voir Figure 2.9).

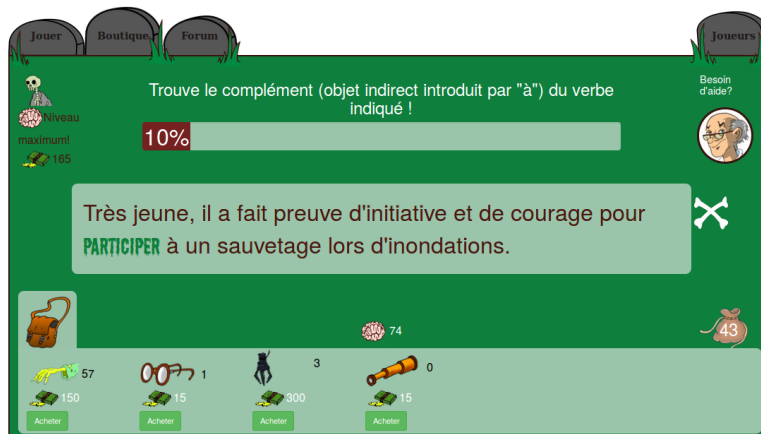


FIGURE 2.9 – Interface principale de jeu dans ZombiLingo : l’élément en surbrillance est la tête de la relation, le joueur doit trouver le dépendant (ici, « à »).

2.2.3 Mécanismes de formation et de contrôle

Bien entendu, ce choix n’est pas sans conséquence, puisque cela induit un biais de confirmation important : les joueurs vont avoir tendance à considérer comme

Chapitre 2. Les jeux ayant un but : un modèle encore incertain

sûr la tête ou le dépendant pré-identifié et ne pas utiliser la croix d'os. Pour contrer ce biais et s'assurer que les consignes d'annotation sont bien comprises par les joueurs, nous avons mis en place deux mécanismes inspirés de **Phrase Detectives**. Le premier est une formation obligatoire sur chaque relation avant de pouvoir la jouer (voir Figure 2.10) avec des retours négatifs en cas d'erreur (voir Figure 2.11).

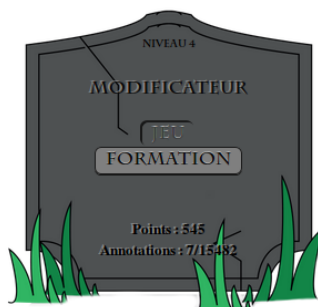


FIGURE 2.10 – ZombiLingo : formation obligatoire avant de pouvoir jouer la relation (ici, modificateur).



FIGURE 2.11 – ZombiLingo : retour au joueur dans le cas d'une erreur lors de la formation, l'erreur est signalée et la solution est indiquée.

Le second est un mécanisme de vérification, déclenché aléatoirement et qui propose une phrase de référence à annoter. Si le joueur donne une réponse fausse, il a un retour négatif comme lors de la formation. Après trois erreurs sur la même relation, il doit refaire la formation. Cette vérification nous sert également à attribuer

un score de confiance au joueur pour une relation que nous allons utiliser ensuite pour sélectionner les meilleures combinaisons d’annotations pour la phrase.

2.2.4 Des résultats très encourageants

Évaluation intrinsèque

Nous avons réalisé une première évaluation, intrinsèque, des résultats obtenus avec ZombiLingo, que nous avons présentée à COLING 2016 (Guillaume et al., 2016). Pour ce faire, nous avons utilisé le corpus Sequoia, que nous avons découpé en deux, une partie servant à la formation et au contrôle des joueurs, l’autre à l’évaluation, comme présenté sur la figure 2.12.

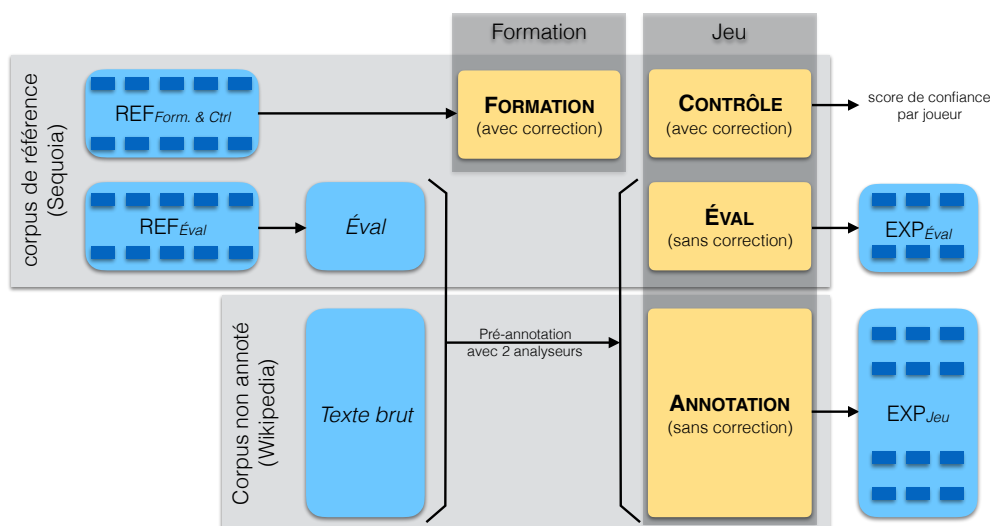


FIGURE 2.12 – Organisation de l’évaluation de ZombiLingo.

À l’époque (juillet 2016), 647 joueurs avaient participé et produit 107 719 annotations sur un total de 128 046 tokens, ce qui représentait le plus gros corpus annoté en syntaxe de dépendances pour le français librement disponible et redistribuable.

En ce qui concerne la qualité des annotations, nous avons calculé la F-mesure des annotations réalisées par les joueurs par rapport aux annotations du corpus Sequoia. Les résultats sont présentés par type de relation dans la figure 2.13 et sont comparés avec ceux des analyseurs syntaxiques Talismane (Urieli, 2013) (à base d’apprentissage) et FrDep-Parse (à base de règles), qui ont été utilisés pour la pré-annotation. ZombiLingo améliore les performances des outils pour une large majorité de relations (toutes sauf quatre), avec une précision moyenne de 0,93.

Pour toutes les relations regroupées dans la partie gauche de la figure, le jeu obtient de meilleurs résultats que les outils. Ces relations ont en commun d’avoir

une densité d’annotation supérieure à 1, autrement dit, elles ont été annotées par plus d’un joueur. Il est donc important pour la qualité des résultats que plusieurs joueurs annotent les mêmes relations.

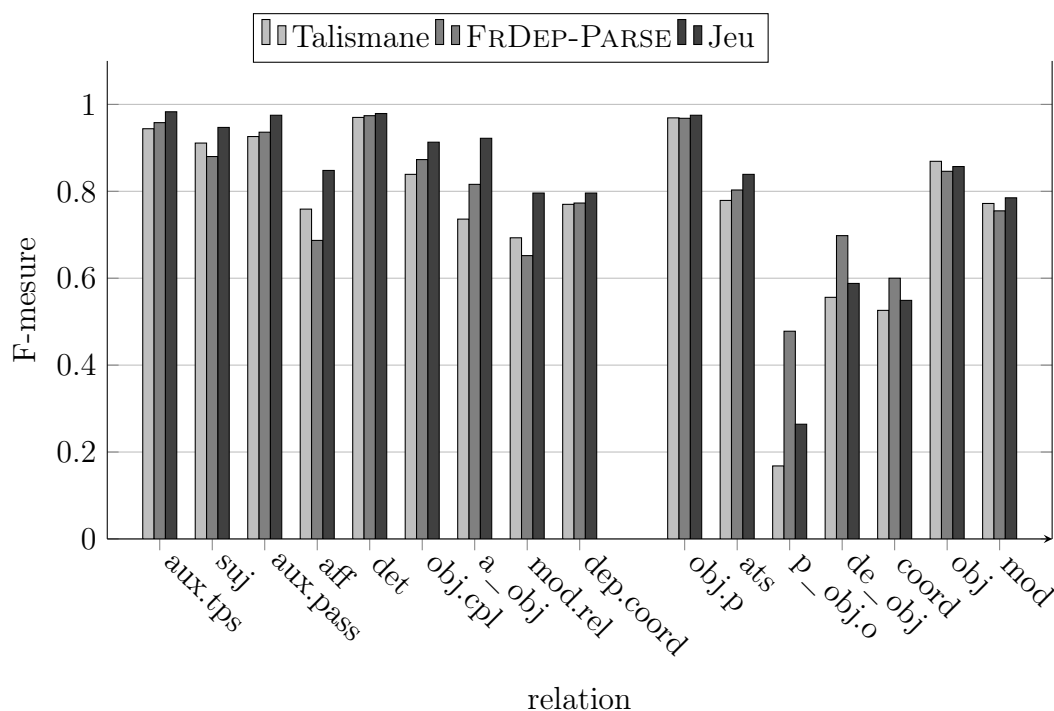


FIGURE 2.13 – F-mesures pour les deux analyseurs syntaxiques utilisés pour la pré-annotation et le jeu, par relation (celles avec une densité supérieure à 1 sont sur la gauche, les autres à droite séparées par une espace).

Nous avons réalisé par la suite une étude portant sur des textes plus spécialisés (sur le thème de l’ADN) dans le cadre de mon cours de Master 2 sur l’annotation collaborative de corpus. Celle-ci a montré des résultats moins clairement positifs, avec une qualité d’annotation globalement comparable à celle du parser Talismane. Cela étant, certaines relations complexes comme **COORD** ont été beaucoup mieux annotées par les joueurs que par l’outil (Fort et al., 2017b).

Évaluation extrinsèque

Diego Alvès, étudiant en Master 1 à la Sorbonne en 2018¹⁰, a réalisé son mémoire sur l’évaluation extrinsèque de ZombiLingo (Alves, 2018), encadré par Bruno Guillaume et moi-même. Dans ce cadre, il a testé l’apport des ressources créées par le jeu sur les performances de deux *parsers* du français, Talismane (Urieli,

10. Diego est aujourd’hui doctorant à l’Université de Zagreb : <https://dfvalio.github.io/>.

2.3 Des jeux aux plateformes (moins) ludifiées

2013) et MaltParser (Nivre et al., 2007). Il a pour ce faire ré-entraîné ces outils en ajoutant au jeu d’entraînement les corpus annotés sur ZombiLingo, puis il a observé les différences de performances sur différentes références.

Les résultats obtenus mériteraient d’être consolidés, mais Diego Alvès a pu montrer que la contribution des joueurs permettrait globalement d’améliorer les performances des outils testés.

Toutefois, et sans surprise, à volume équivalent de données, le gain serait plus important lorsque l’on utilise le corpus arboré de Paris 7 (Abeillé et al., 2003) (dit « French Treebank »), qui a été annoté par des experts en syntaxe.

2.2.5 De ZombiLingo à ZombiLUDik

La ressource a continué à évoluer dans les années qui ont suivi, au fil des ajouts de corpus sur la plateforme ZombiLingo. Ainsi, en février 2022, on comptait 1 591 inscrits sur la plateforme, qui avait produit 504 178 annotations. Depuis, le nom de domaine a été résilié, mais le site reste disponible sur <http://gwap.grew.fr/>.

À partir de fin 2018, nous avons décidé de prendre le temps d’adapter le jeu au formalisme du projet Universal Dependencies (UD)¹¹. Bruno Guillaume a donc créé un clone de ZombiLingo pour UD, ZombiLUDik¹². Notre idée était de proposer cet outil pour faciliter l’annotation en syntaxe de dépendances de type UD pour les langues les moins dotées. Afin d’en montrer la facilité d’utilisation, nous l’avons d’ailleurs adapté pour l’anglais.

ZombiLUDik est en ligne depuis mai 2019. Il a attiré à ce jour 139 inscrits, qui ont produit 53 469 annotations. Nous avons souhaité nous concentrer sur l’annotation de français oral et j’ai mobilisé sur ce sujet un étudiant de Master 1 dans le cadre de son mémoire, avec l’aide de mes Master 2. Malheureusement, l’engagement des étudiants a été un peu décevant et la pandémie n’a pas permis que nous continuions.

2.3 Des jeux aux plateformes (moins) ludifiées

La conception et le développement d’un jeu ayant un but sont très chronophages. Ainsi, il nous a fallu près d’un an pour finaliser une première version stable de ZombiLingo, comprenant un espace réservé à son administration (*backend*). Nous avons heureusement pu bénéficier d’un financement Inria (dit ADT) qui nous a permis de recruter un ingénieur, Nicolas Lefèbvre, pendant deux ans. Il m’a par conséquent semblé intéressant de rentabiliser cet investissement en ré-

11. Voir : <https://universaldependencies.org/>.

12. Voir : <https://zombiludik.org/>.

utilisant des composants pour créer une nouvelle plateforme d’une part et de tester une ludification légère, plus facile à concevoir, d’autre part.

2.3.1 Rigor Mortis : annoter les unités polylexicales

Rigor Mortis¹³ est issu du besoin d’annoter les unités polylexicales, notamment dans UD. La plateforme est donc une extension de ZombiLingo, revendiquée dans la proximité de la thématique puisque *Rigor mortis* est l’expression latine pour « rigidité cadavérique ». Cette rigidité fait bien sûr référence au figement des expressions à identifier.

La plateforme a été réalisée dans le cadre du mémoire de Master 1 de Yann-Alan Pilatte, avec l’aide de Nicolas Lefèbvre et en collaboration avec Bruno Guillaume et Mathieu Constant (ATILF).

Une plateforme ludifiée

La plateforme comprend trois parties différentes, qui sont débloquentes séquentiellement. La ludification de la plateforme est limitée à un design lié à l’univers des momies et des pyramides, au déblocage des niveaux et à un classement des joueurs en fonction de leur production (voir Figure 2.14).



FIGURE 2.14 – Interface principale de Rigor Mortis.

La première phase du jeu a été pensée et conçue entièrement par Yann-Alan Pilatte et correspond à un test de l’intuition des locuteurs concernant les unités polylexicales : sont-ils capables, sans formation aucune, de repérer les unités polylexicales d’un texte ?

La deuxième phase consiste en une formation des participants à ce qu’est une unité polylexicale. Pour cela, nous utilisons certains des critères définis dans le

13. Voir : <http://rigor-mortis.org/>.

2.3 Des jeux aux plateformes (moins) ludifiées

cadre du projet PARSEME-FR¹⁴ : test de remplacement (« cordon bleu » *vs* « cordon rouge »), test « cranberry » (un mot qui n'existe pas en dehors de l'expression, comme « perlimpinpin »), test d'insertion (« Luc prend la puissante mouche » n'est pas possible), test morphosyntaxique (« ramener ses fraises » n'a pas le même sens que « ramener sa fraise ») et test de déterminant nul (on ne peut pas « prêter la main forte »). Si l'un de ces tests est applicable, alors il s'agit d'une unité polylexicale. Nous les avons choisis car ils sont les plus productifs tout en étant relativement faciles à comprendre. La phase deux est la seule pendant laquelle les participants ont des retours immédiats sur leur annotation (voir Figure 2.15).



FIGURE 2.15 – Phase deux de Rigor Mortis : formation des participants.

Enfin, la troisième phase est la phase de production d'annotations en tant que telle, lors de laquelle les participants identifient des unités polylexicales dans des phrases extraites d'un corpus.

Intuition *vs* formation

Nous avons évalué séparément la partie intuition de la partie annotation après formation.

Sur la première, nous avons eu entre 65 et 68 participants qui ont annoté les dix phrases de test d'intuition, avec un rappel de 65,05 % sur les unités polylexicales non fonctionnelles (par exemple, « dommages et intérêts »), celles qui

14. Voir : <https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Criteres>.

sont les plus productives (Fort et al., 2018a). Pour les unités polylexicales fonctionnelles (par exemple, « entre autre »), les performances s’effondrent à 30,41 %. Dans notre expérience d’annotation, ces expressions sont tellement courantes que nous ne les voyons plus. D’ailleurs, ce sont généralement les premières unités polylexicales identifiées dans une langue. Cette expérience, bien que limitée par le nombre de phrases annotées, montre que l’intuition des locuteurs concernant les unités polylexicales est d’un niveau satisfaisant sur les unités non fonctionnelles.

Une fois la phase trois débloquée, les participants sont considérés comme formés et il est alors possible d’évaluer l’impact de la formation sur la qualité de leur production. Pour ce faire, nous avons établi une référence *a posteriori*. En effet, il s’est avéré que 62 phrases de la phase trois ont été annotées sur la plateforme par au moins deux membres du projet PARSEME-FR¹⁵, dont 32 par trois d’entre eux (Carlos Ramish, Agata Savary et Mathieu Constant). Nous sommes repartis de leurs annotations et les avons corrigées lorsqu’elles divergeaient (adjudication), afin de produire une référence (Fort et al., 2020).

Chaque phrase de référence a été jouée par une moyenne de 31,48 joueurs (entre 22 et 51). Nous avons établi empiriquement un seuil optimal du nombre minimal de joueurs en accord sur une annotation pour obtenir le meilleur résultat (voir Figure 2.16). La F-mesure maximale, 0,685, est obtenue à un seuil de 25 %, mais à 36 % on atteint un équilibre entre précision et rappel (avec une F-mesure à 0,622).

Les phrases étant plus complexes que celles de la première phase, ce résultat tend à prouver l’impact positif de la formation proposée en phase deux. Par ailleurs, le pourcentage d’accord sur chaque annotation peut être utilisé pour mesurer le continuum du figement lexical.

2.3.2 Bisame et Recettes de grammaire : appliquer la myriadisation aux langues non-standardisées

Alice Millour a commencé son doctorat avec moi à la Sorbonne¹⁶ en 2016. Elle m’avait contactée l’année précédente pour son Master 2 Recherche, car elle souhaitait faire de la recherche en TAL sur les langues peu dotées et savait que le sujet m’intéressait. Nous avons donc commencé à travailler ensemble à partir de fin 2015 sur la myriadisation pour les langues peu dotées, en particulier non standardisées (sans orthographe normée). Elle a soutenu sa thèse en décembre 2020 (Millour, 2020).

15. Cela n’est pas surprenant, puisque nous avons communiqué sur cette plateforme sur les listes de diffusion du domaine.

16. Son encadrant HDR était Claude Montacié, qui m’a fait confiance et m’a très largement laissée encadrer Alice (à 90 %).

2.3 Des jeux aux plateformes (moins) ludifiées

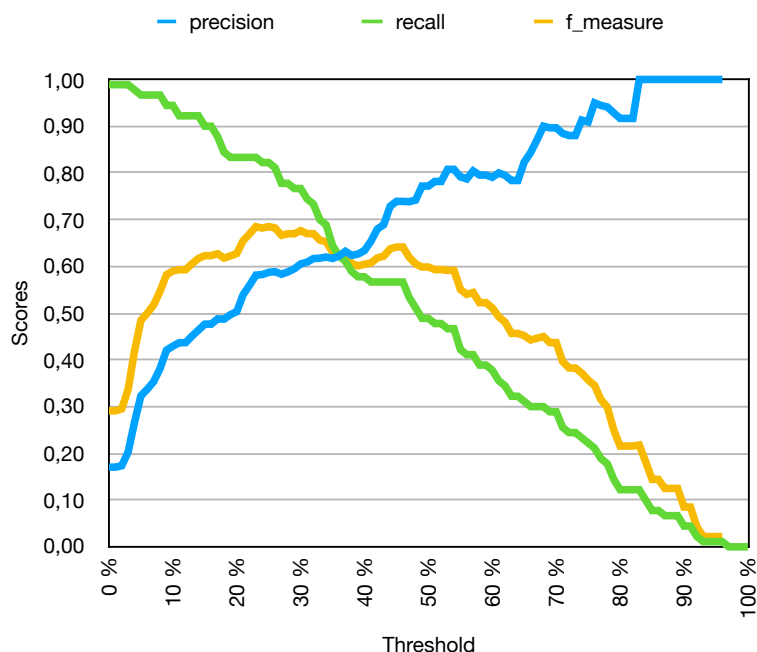


FIGURE 2.16 – Figure 8 de (Fort et al., 2020) : précision, rappel et F-mesure en fonction du seuil d’accord.

Bisame : myriadiser les parties du discours

Du fait de contraintes internes à la Sorbonne, nous ne pouvions travailler que sur des langues de France. Delphine Bernhard (LiLPa / Université de Strasbourg) nous ayant assuré de son intérêt pour le projet, nous avons décidé de commencer par l’alsacien, tout en ayant pour but de créer des méthodologies applicables pour toutes les langues. Pendant son M2, Alice Millour a donc développé Bisame (Millour et al., 2017; Millour and Fort, 2018c)¹⁷, une plateforme légèrement ludifiée pour l’annotation en parties du discours. Cette plateforme inclut une phase de formation des participants et un peu de ludification sous la forme d’un classement (*leaderboard*).

Parmi les 180 personnes qui se sont inscrites sur le site, 42 ont réellement participé et ont produit 15 846 annotations. Afin d’alléger la tâche, l’interface d’annotation présente en priorité aux participants une ou deux étiquettes (voir Figure 2.17) résultant d’une pré-annotation avec deux outils différents, TreeTagger

17. La plateforme était originellement ici, mais une redirection vers Recettes de grammaire a été ajoutée : <http://bisame.paris-sorbonne.fr/>.



FIGURE 2.17 – Annotation directe.

adapté pour l’alsacien (Bernhard and Ligozat, 2013) et MElt (Denis and Sagot, 2010) entraîné avec le corpus de référence. Des phrases de référence sont insérées parmi les phrases à annoter et servent à calculer dynamiquement un score de confiance pour le participant et pour son annotation. Ce dernier sera ensuite utilisé pour calculer l’étiquette la plus probable parmi celles proposées par les participants.

Au total, 7 750 annotations ont été produites sur notre corpus de référence de 1 468 tokens. Ces annotations sont d’une qualité tout à fait satisfaisante, avec une moyenne des F-mesures pondérées par les effectifs de 0,93.

Alice Millour a ensuite entraîné un *tagger* MElt avec les annotations produites par la plateforme. En y ajoutant un lexique, nous obtenons une exactitude de 82 %, ce qui correspond à peu près aux performances obtenues par Bernhard and Ligozat (2013) à l’époque, avec la possibilité de produire davantage d’annotations et d’améliorer ainsi le *tagger*, ce qui est impossible avec la méthode proposée par nos collègues.

Dans le cadre du travail de M1 de Gwladys Feler¹⁸, une étudiante locutrice du créole guadeloupéen qu’elle a co-encadrée avec moi, Alice Millour a adapté la plateforme pour cette langue. Si l’expérience a été positive du point de vue de l’adaptabilité de la plateforme, elle n’a cependant pas été concluante concernant les résultats obtenus. Nous avons en effet eu très peu de participation (11 participants effectifs) et la qualité n’a pas été à la hauteur de nos espérances, principalement du fait du manque d’annotations (seulement 1 205) (Millour and Fort, 2018b). De fait, la myriadisation bénévole d’une nouvelle langue nécessite du temps (de prise de contact avec la communauté des locuteurs, en particulier) et s’accorde mal avec les délais d’un mémoire de M1.

18. Cette étudiante nous a rendu son mémoire mais ne l’a malheureusement jamais soutenu, elle voulait changer d’orientation et n’a pas souhaité aller au bout du travail.

2.3 Des jeux aux plateformes (moins) ludifiées

Recettes de grammaire : produire des données pour les langues non-standardisées

Encouragées par les résultats obtenus sur l’alsacien, mais conscientes du manque de corpus bruts à annoter, nous avons proposé et obtenu le financement du projet PLURAL (Production LUDique de Ressources Annotées pour les Langues de France) dans le cadre de l’appel « Langues et numérique » du Ministère de la culture en 2018. Ce projet était une collaboration avec Bruno Guillaume, Delphine Bernhard et André Thibault (professeur en sciences du langage, spécialiste de la variation, à Sorbonne Université).

Alice Millour a ainsi pu avoir l’aide de Nicolas Lefèbvre pour développer une autre plateforme de myriadisation, *Recettes de grammaire*¹⁹. Cette plateforme avait pour but de collecter des données pour les langues non-standardisées : des corpus de textes bruts, des annotations en parties du discours et des variantes d’unités lexicales. Nous avons eu l’idée de proposer pour cela aux participants de partager des recettes (voir Figure 2.18). Nous avons ensuite élargi les possibilités à toutes sortes de textes, notamment des poèmes et des proverbes. La plateforme a d’abord été instanciée pour l’alsacien, puis pour le créole mauricien, dans le cadre du mémoire de M1 d’une étudiante locutrice, Harmonie Begue.



FIGURE 2.18 – Bandeau d’accueil de *Recettes de grammaire* pour l’alsacien.

L’interface est inspirée de celles de sites de partage de recettes et propose des fonctionnalités du type « Like » ou ajout de commentaires. Elle est un peu ludifiée, avec des avatars, des points à gagner (voir Figure 2.18 à droite) et des badges (voir Figure 2.19).

Malgré une interface plus attrayante, la plateforme a attiré moins de participants (seulement 55 inscrits pour l’alsacien et 17 pour le mauricien) que *Bisame*. Pour l’alsacien, seuls sept participants ont proposé des textes, pour un total de

19. Le site n’est plus accessible du fait de problèmes de certificat : <http://bisame.paris-sorbonne.fr/recettes/>

Chapitre 2. Les jeux ayant un but : un modèle encore incertain



FIGURE 2.19 – Badges que les participants peuvent obtenir dans *Recettes de grammaire* pour l’alsacien, en fonction de leur activité sur les différentes tâches.

1 803 tokens et 12 ont produit 200 annotations. Pour le créole mauricien, 12 inscrits ont ajouté 36 textes pour un total de 1 903 tokens et sept ont produit 1 050 annotations.

La fonctionnalité de collecte de variantes, la seule du genre à ma connaissance, a permis de récolter 215 variantes de 106 mots de l’alsacien et 46 sur 36 mots pour le créole mauricien (voir Figure 2.20).

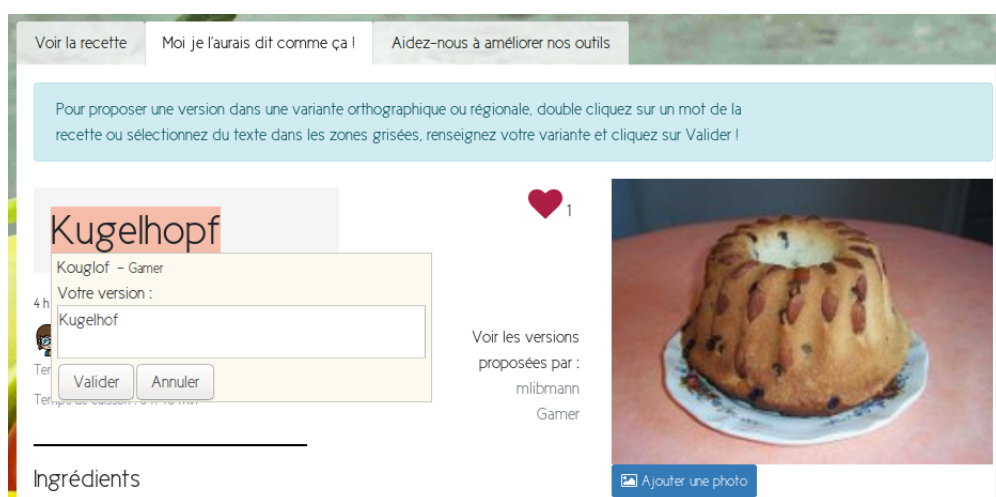


FIGURE 2.20 – Ajout de variantes dans *Recettes de grammaire* (alsacien).

Nous avons utilisé les variantes de l’alsacien pour en générer d’autres et ainsi améliorer la robustesse du modèle de tagger aux variations dialectales et scriptu-

rales de la langue (Millour and Fort, 2019). Un expert seul n’aurait pas pu produire ces variantes, seule la participation active des locuteurs permet de capter une partie de la diversité de la langue. De ce point de vue, la myriadisation est un outil précieux. Par ailleurs, la méthodologie que nous avons mise en place est applicable à n’importe quelle langue non-standardisée.

Cependant, la plateforme n’a pas tenu ses promesses concernant la collecte de corpus bruts, en particulier sur l’alsacien, langue sur laquelle nous avons de l’expérience et des participants (ceux de Bisame). Alice Millour a donc mené des enquêtes auprès des locuteurs de l’alsacien (Millour, 2019) et du créole mauricien (avec Harmonie Begue) (Begue, 2019) sur l’utilisation de leur langue en ligne pour essayer de comprendre les raisons de ce relatif échec. Il en ressort que si les locuteurs écrivent leur langue (environ 70 % des répondants), seuls 46 % des répondants pour l’alsacien évaluent leur expression écrite comme bonne ou moyenne (72 % pour le créole mauricien). Cette impression d’incompétence (évidemment injustifiée) pourrait expliquer leur réticence à écrire en alsacien. Le fait que la plateforme était hébergée sur un serveur de la Sorbonne a pu malheureusement amplifier ce phénomène.

Nous aurions sans doute dû réaliser cette enquête avant de concevoir la plateforme, que nous aurions alors rendu plus pédagogique et rassurante pour les locuteurs. La gravité du manque de confiance en eux de ceux-ci nous a surprises et la prise de conscience est arrivée trop tard dans la thèse pour que nous puissions véritablement corriger le tir. Ce relatif échec ne remet cependant pas en cause l’intérêt des résultats obtenus, mais montre la difficulté de l’entreprise poursuivie.

Alice Millour a en effet mené pendant sa thèse une recherche exigeante, sur un sujet qui pousse le TAL dans ses retranchements, puisqu’ancré dans la variation et le manque de données. Elle a su, pour avancer, sortir du confort disciplinaire pour se saisir d’outils de la linguistique de terrain, de la sociolinguistique et des sciences de l’information et de la communication. Cette première expérience d’encadrement de thèse a été pour moi très riche, au niveau scientifique comme humain. Mes encadrements actuels lui doivent beaucoup.

2.4 Les incertitudes du modèle

Ces derniers exemples montrent à quel point le succès d’une plateforme de myriadisation ludifiée dépend de facteurs variés, difficiles à prévoir. Outre les limites liées aux participants eux-mêmes (par exemple, l’insécurité des locuteurs de l’alsacien), la conception d’un jeu n’a rien d’une recette de cuisine. Bien entendu, s’il existe des formations en conception de jeux, c’est bien que certaines règles ont été identifiées. Cependant, aucun cours ne peut remplacer la capacité à trouver

la bonne idée. Ainsi, le succès du jeu 2048²⁰ est difficile à justifier, en particulier étant donné le faible taux de victoire (environ 1 % des parties jouées, selon Wikipédia²¹). Il faut ajouter à cela le fait qu'un jeu ayant un but n'est pas un jeu comme les autres, car il est créé avant tout pour produire des données, et ces données doivent être de qualité.

2.4.1 Créer un jeu, un savoir-faire complexe

Concevoir, développer et maintenir un jeu prend du temps. Si nous avons pu développer un prototype de ZombiLingo assez rapidement (environ six mois), il nous a fallu plus d'un an pour obtenir une version réellement déployable. Cela coûte donc cher et il faut que l'investissement soit « rentable », autrement dit que la quantité et la qualité des ressources produites soient à la hauteur.

Pour assurer une production de données massive, il faut attirer des joueurs, il faut donc les connaître, savoir ce qui les intéresse et pourrait les pousser à venir sur la plateforme (motivation) puis y revenir (volition (Fenouillet et al., 2009)).

Mathieu Lafourcade, pionnier du domaine en France, a par exemple utilisé pour créer son jeu une grille utilisée par la CIA (Central Intelligence Agency) pour recruter des sources²², MICE²³, pour *Money* (argent ou récompense), *Ideology* (idéologie ou intérêt), *Constraint* (contrainte) et *Ego* (place dans la communauté).

Nous nous sommes intéressés au sujet et parmi les nombreuses typologies disponibles, nous avons jeté notre dévolu sur la plus connue, proposée par Richard Bartle (Bartle, 1996). Selon lui, les joueurs ont quatre principales sources d'intérêt dans un jeu : le gain en points et en niveaux (*achievement*), l'exploration des arcanes du jeu (*exploration*), les interactions avec les autres (*socialising*) et le pouvoir sur les autres (*killing*). Il organise ces centres d'intérêt selon deux axes, représentant d'une part les actions et interactions, et d'autre part le monde et les joueurs (voir Figure 2.21). Ainsi, les *Achievers* aiment agir sur le monde, alors que les *Explorers* veulent interagir avec lui. Quant aux *Killers*, ils agissent sur les autres joueurs, alors que les *Socialisers* interagissent avec eux. Bien entendu, il ne s'agit pas d'une échelle binaire : un *Achiever* peut avoir des envies d'exploration, mais ce ne sera pas son moteur principal.

Un jeu, pour attirer le plus de participants possible, devrait donc proposer des fonctionnalités satisfaisant différents profils de joueurs. Nous avons essayé de prendre cela en compte dans ZombiLingo, en proposant, outre le classement (*lea-*

20. Voir par exemple : <https://jeu2048.fr/>.

21. Voir : [https://fr.wikipedia.org/wiki/2048_\(jeu_vid%C3%A9o\)](https://fr.wikipedia.org/wiki/2048_(jeu_vid%C3%A9o)).

22. Communication personnelle, renouvelée le 7 avril 2022.

23. Voir : <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol.-57-no.-1-a/vol.-57-no.-1-a-pdfs/Burkett-MICE%20to%20RASCALS.pdf>.

derboard) et les badges pour les *Achievers*, des possibilités d'exploration pour les *Explorers* (notamment un accès à un autre jeu caché dans une coccinelle, qui passe parfois dans le jeu), un forum pour les *Socialisers* et des duels pour les *Killers*²⁴.

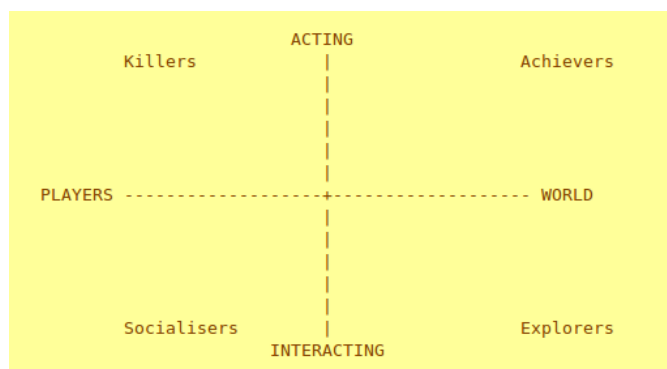


FIGURE 2.21 – Typologie des joueurs selon (Bartle, 1996).

Nous avons moins fait d'efforts en ce sens pour *Rigor Mortis*, du fait du manque de temps et de l'absence de développeur dédié.

2.4.2 Des contraintes des jeux ayant un but

Comme je le rappelle plus haut, les jeux ayant un but ne sont pas de jeux comme les autres. En effet, la production de données de qualité impose des contraintes sur le jeu à créer qu'il ne faut jamais oublier lors du développement : un joueur ne doit gagner des points que s'il produit des données de qualité et la production de données de qualité doit permettre de gagner des points sans pour autant impacter l'intérêt du jeu. Or, préserver ce cercle vertueux n'est pas si simple.

Ainsi, nous avons prévu dans *ZombiLingo* que les phrases présentées au joueur pouvaient parfois s'effacer peu à peu ou rapetisser, de manière aléatoire. Seule l'utilisation d'un objet dédié (une longue vue ou une paire de lunettes) pouvait contrecarrer cet effet. Cette fonctionnalité nous plaisait beaucoup, car elle permettait de surprendre le joueur, donc de le « réveiller » s'il avait tendance à être pris dans une routine de jeu et de relancer son intérêt. De fait, l'effet de surprise jouait à plein, à tel point que les joueurs se mettaient à paniquer et à cliquer n'importe où dans l'urgence, ce qui générerait de mauvaises annotations. Bien entendu, dès que nous nous en sommes rendus compte, nous avons désactivé le clic de souris à l'enclenchement de l'événement, mais cela montre qu'il faut être très attentifs aux effets potentiels d'une fonctionnalité sur les données.

24. Un duel peut aussi être considéré comme une interaction, donc plutôt de l'ordre de la socialisation, mais il peut aussi ralentir l'autre joueur.

La situation inverse est également possible. Un joueur a ainsi trouvé une faille dans le code de **JeuxDeMots** pour avoir plus de temps de jeu. De ce fait, il a pu créer plus de données, de bonne qualité. Cette tricherie était une bonne chose pour la production, mais avait un impact négatif sur le jeu, car les autres joueurs s'en sont rendus compte et menaçaient de partir. Mathieu Lafourcade a donc dû bannir le joueur, alors qu'il produisait de très bonnes données ²⁵.

Au-delà de ces deux cas d'école, créer une interface vraiment ludique qui permette de créer des données de qualité nécessite beaucoup d'énergie, du temps et des compétences variées. Il faut ajouter à cela des compétences en communication auprès du grand public que nous ne possédons pas forcément ²⁶, mais qui sont indispensables pour attirer des participants.

2.4.3 Motiver les participants

Nous avons déployé beaucoup d'énergie à faire la publicité de nos plateformes, en particulier de **ZombiLingo**, sur les listes de diffusion du domaine, les réseaux sociaux, la presse grand public, les événements comme la Fête de la science, les forums et les publications de vulgarisation. Bien que nous n'ayons pas fait d'analyse précise de l'impact de chacun de ces media, nous avons clairement vu l'impact de certaines publications, en particulier dans la presse grand public ²⁷, sur les inscriptions. Cela étant, ces nouveaux joueurs ne sont en général pas revenus et n'ont produit que peu de données. Nous avons donc mené une enquête auprès des joueurs de **ZombiLingo** pour mieux les connaître et comprendre leurs motivations ([Fort et al., 2017a](#)).

Pour cela, nous avons contacté par courriel ceux qui avaient renseigné ce champ lors de leur inscriptions (un champ facultatif, pour des raisons de minimisation de données) en leur proposant un questionnaire. Nous avons obtenu 109 réponses sur les 996 joueurs inscrits à l'époque, dont 515 avaient laissé leur adresse. Vingt d'entre eux étaient de gros joueurs (ayant produit plus de 500 annotations ou ayant joué plus de cinq jours différents).

Si la participation sur le jeu est équilibrée entre hommes et femmes, les gros joueurs sont majoritairement des hommes (65 %). À titre de comparaison sur le sujet, **Phrase Detectives** a quant à lui attiré 65 % de femmes et **JeuxDeMots** 60 % (sans distinction entre gros et petits joueurs) ([Chamberlain et al., 2013](#)). Nos gros joueurs ont en général moins de 40 ans et un haut niveau d'études (75 % d'entre eux ont au moins un niveau Master). Parmi les autres joueurs, 57 % ont au

25. Communication personnelle renouvelée le 7 avril 2022.

26. Je tiens ici à remercier les services communication de la Sorbonne et du LORIA, qui nous ont aidés à faire la publicité de nos plateformes.

27. Pour **ZombiLingo**, nous avons eu un article dans l'Est Républicain en 2015 et un autre dans Sciences et Avenir, en 2017.

moins un niveau Master et 32 % un niveau licence. Enfin, 25 % des gros joueurs viennent du domaine du TAL et 15 % de la linguistique, alors que 74 % des autres joueurs sont totalement extérieurs à ces disciplines.

Les réponses concernant leurs motivations sont particulièrement intéressantes, puisque les gros joueurs citent comme motivations principales le jeu, la linguistique et l'aide à la science (chaque réponse étant sélectionnée par 55 % d'entre eux), alors que les autres joueurs mettent en avant la curiosité (70 %) et l'aide à la science (66 %). Cela confirme que la ludification est un facteur important pour attirer de gros joueurs et accroître la quantité de données produite.

Cela étant, l'expérience montre qu'il faut tout de même relancer régulièrement la participation, même des gros joueurs, par le biais de nouveautés, comme des défis, des nouveaux corpus, de nouvelles fonctionnalités.

Tuite (2014) plaide pour des jeux ayant un but qui permettent aux participants de comprendre à quoi ils participent, d'apprendre quelque chose, des jeux, en somme, qui permettent une encapacitation (*empowerment*) des joueurs. Je suis particulièrement sensible à ces arguments depuis mes échanges avec mes collègues du Muséum national d'histoire naturelle sur les sciences participatives. Il me semble en effet que si nous parvenions à aligner les besoins de la recherche avec ceux des locuteurs (ce qui reviendrait à passer d'une motivation extrinsèque à une motivation intrinsèque), nous n'aurions plus à mettre autant d'énergie dans la publicité et nous pourrions recréer du lien entre citoyens et chercheurs.

Vue du prisme des langues non-standardisées, il m'a semblé qu'une instanciation de cette idée pourrait être de favoriser la transmission inter-générationnelle. En effet, la plupart des langues non-standardisées souffrent de la rupture de la transmission de celles-ci entre les générations. Afin de favoriser les échanges entre enfants et grands-parents (ou autres personnes de la famille maîtrisant la langue), j'ai donc imaginé qu'on pourrait concevoir un jeu de quête où les jeunes joueurs auraient besoin de l'aide de leurs aînés pour résoudre les énigmes et remporter le prix.

Cette idée assez vague est à l'origine du jeu *Katana et Grand Guru*²⁸ (voir Figure 2.22), que nous avons développé lors de deux hackathons organisés dans le cadre de l'action COST enetCollect. Notre équipe était composée de ma doctorante, Alice Millour, d'un de mes étudiants de M2, Yann-Alan Pilatte, d'une professeure des écoles, Marianne Araneta, et de deux chercheuses en linguistique, Ivana Lazić Konjik et Annalisa Raffone. *Katana* est un jeu de rôle à l'ancienne (en deux dimensions), dont le héros doit sauver un monde dans lequel les mots ont perdu leur pouvoir, lui seul ayant encore celui de faire vivre les choses qu'il nomme (voir Figure 2.23).

Le scénario est décomposé en quatre parties ou niveaux, représentant quatre

28. Voir : <https://bisame.paris-sorbonne.fr/lost-words/index.html>.

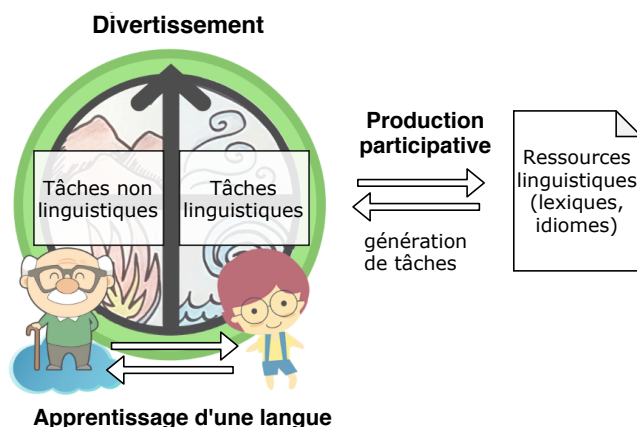


FIGURE 2.22 – Katana et Grand Guru : les interactions entre le jeune joueur et son aîné permettent à la fois de produire des données langagières et de favoriser la transmission de la langue entre générations.

villages qui correspondent chacun à un élément (terre, eau, feu, air). Seul le premier niveau, celui de la terre, est implémenté (voir son scénario dans la figure 2.24) et il est instancié pour l'irlandais. Nous avons établi les caractéristiques des personnages et du village pour chaque niveau, mais n'avons pas eu le temps de les développer. La collecte de données concerne ici uniquement du lexique simple, y compris, bien entendu, des variantes dialectales et scripturales.



FIGURE 2.23 – Katana et Grand Guru : le mot irlandais pour arbre (« crann ») permet de faire reverdir les arbres, mais en cas d'erreur les arbres restent morts.

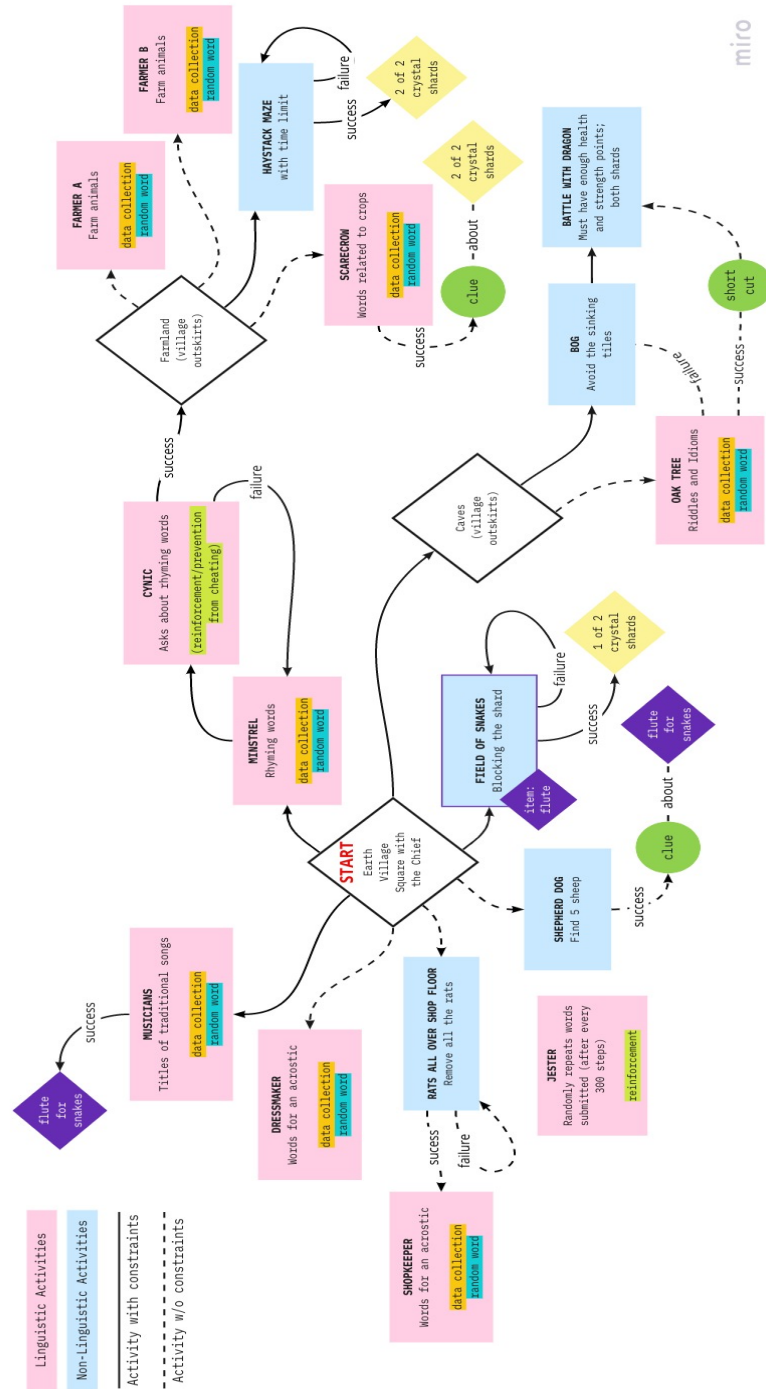


FIGURE 2.24 – Katana et Grand Guru : scénario du premier niveau.

Nous avons mis en place des mécanismes de vérification avec des mots de référence pour s'assurer d'une part que le joueur ne saisit pas n'importe quoi (voir Figure 2.23), et d'autre part, qu'il assimile le vocabulaire appris (Millour et al., 2019).

Nous avons pour projet de tester le jeu en deux temps : d'abord de le traduire en français et de l'instancier avec l'anglais, pour en tester la jouabilité parmi des élèves francophones apprenant l'anglais, puis de l'évaluer en situation réelle dans une communauté linguistique. Nous avons pris contact avec Josh Holden, un collègue linguiste de l'Université de Blue Quills, qui travaille sur le déné, une langue athapascane, et qui était très intéressé, mais la pandémie de COVID a mis en pause nos efforts pour l'instant. Le jeu est cependant traduit en français et nous envisageons de le tester avec Alice Millour auprès d'enfants francophones. *Katana* est une très belle production et j'espère pouvoir la mettre en valeur dans les années à venir.

Malgré les incertitudes liées aux jeux ayant un but, je reste convaincue de leur puissance, non seulement en tant que moyen de production de données de qualité, mais également comme levier d'encapacitation des communautés de locuteurs. Ils représentent une alternative éthique crédible aux plateformes de micro-travail.

Les problèmes éthiques du TAL ne se limitent cependant pas à la production de données et j'ai élargi mes recherches à l'éthique dans et pour le TAL en général depuis ma prise de poste à la Sorbonne en 2014.

L'éthique dans et pour le TAL

Sommaire

3.1	Créer le chemin vers la légitimité	55
3.2	Une éthique morcelée par la spécialisation	63
3.3	Pour une vision plus systémique de l'éthique dans le TAL	71

3.1 Créer le chemin vers la légitimité

3.1.1 Des questions invisibles en TAL avant 2011

J'ai commencé à m'intéresser à l'éthique dans le TAL en 2010, par le prisme des problèmes que posait (et pose encore, dans une large mesure) l'utilisation de la plateforme de travail parcellisé **Amazon Mechanical Turk**. Ce questionnement a donné lieu à un article *Last Words* dans la revue *Computational Linguistics* avec Gilles Adda (LISN) et Kevin Bretonnel Cohen (University of Colorado, School of Medicine) ([Fort et al., 2011](#)). Cet article a été beaucoup cité et a permis d'ouvrir les yeux de certains collègues sur la réalité du travail sur cette plateforme¹. C'était notre but en l'écrivant. À ma connaissance, les premiers à avoir soulevé ces questions sont cependant Gilles Adda et Joseph Mariani (retraité, à l'époque LIMSI), dans un article présenté dans un atelier de LREC 2010 ([Adda and Mariani, 2010](#)).

Avant eux et ces questions, il est difficile de trouver des publications dans des conférences ou revues de TAL ayant trait à l'éthique. En 2010, un article très complet a été publié sur l'éthique de la recherche utilisant les données de Facebook ([Zimmer, 2010](#)), mais en technologies de l'information, et en 2011 un autre sur l'éthique de la traduction automatique ([Kenny, 2011](#)), mais dans une conférence de traductologie. Les quelques références à l'éthique que l'on trouve dans des publications du TAL sont d'ailleurs liées aux ressources utilisées pour

1. C'est par exemple le cas pour Chris Callison-Burch (University of Pennsylvania), qui a changé sa manière de travailler sur la plateforme suite à la lecture de l'article [Communication personnelle, 24 mars 2021].

la traduction automatique par l'exemple (Drugan and Babych, 2010). À l'époque, en dehors de quelques articles isolés concernant les données utilisées, les questions éthiques que pose le TAL sont donc encore très largement invisibles. Si l'on regarde cet état de fait sous l'angle de l'impact du domaine sur l'humain, il prend tout son sens : le seul rapport qu'entretenait le TAL avec des être humains (en dehors des membres de la discipline) était celui d'un consommateur de données avec ses fournisseurs. Les applications du TAL dans la vie réelle étaient à l'époque encore trop peu performantes, donc trop peu utilisées, ou trop peu visibles pour poser problème. À l'inverse, le domaine frère du traitement automatique de la parole a dû se positionner beaucoup plus tôt sur les dangers de l'utilisation de la reconnaissance de la parole dans les tribunaux (Bonastre et al., 2003; Bonastre, 2020).

À partir de 2014, à ma prise de poste à la Sorbonne, je me suis efforcée de faire bouger les lignes en créant des espaces de réflexion au niveau national. Je présente ces activités dans la figure 3.1.

3.1.2 Des espaces pour réfléchir, ensemble

À l'automne 2014, j'ai organisé avec Benoît Sagot (Inria Paris) et avec le soutien appuyé de Patrick Paroubek (LISN) alors président de l'ATALA, le premier événement sur le sujet, une journée d'études ATALA intitulée « éthique et TAL »². Le succès de cette journée, à la fois en termes de nombre de présentations (huit) et de participants (jusqu'à soixante) m'a encouragée à organiser, avec Maxime Amblard (Université de Lorraine / LORIA) et Gilles Adda, un atelier sur le sujet dans le cadre de la conférence TALN 2015, ETeRNAL (Ethique et TRaitemeNt Automatique des Langues)³, reconduit en 2020⁴. J'ai profité de la dynamique ainsi créée pour proposer un numéro spécial de la revue TAL en 2016, dont j'ai été la rédactrice en cheffe invitée avec Gilles Adda et Kevin Bretonnel Cohen⁵.

Ces espaces successifs ont permis de porter des réflexions sur des sujets aussi variés que l'anonymisation (Amblard et al., 2014; Eshkol-Taravella et al., 2014; De Mazancourt et al., 2014), la diversité linguistique (Enguehard and Mangeot, 2014), l'*open source* (De Chalendar, 2014), les jeux ayant un but (Lafourcade and Lebrun, 2014), l'évaluation (Mathet and Widlöcher, 2016; Garnerin et al., 2020; Lion-Bouton et al., 2020), les conséquences de l'utilisation réelle des applications de traitement des langues et de la parole (Antoine and Lefeuvre, 2014; Lefeuvre-Halftermeyer et al., 2016; Bonastre, 2020), l'utilisation de données issues du Web (Barbaresi and Lejeune, 2020), la répliquabilité (Millour et al., 2020) ou l'éthique du TAL en général (Amblard, 2016). Nous avons donc vu abordés en

2. Voir : https://www.schplaf.org/kf/JE_ATALA.html.

3. Voir : <http://talnarchives.atala.org/ateliers/2015/ETeRNAL/index.html>.

4. Voir : <http://talnarchives.atala.org/ateliers/2020/ETeRNAL/index.html>.

5. Voir : <https://www.atala.org/content/tal-et-%C3%A9thique>.

3.1 Créer le chemin vers la légitimité

France, dès avant 2017, des sujets qui sont entre temps devenus, sinon à la mode, du moins plus couramment traités (voir Section 3.2). La tâche n’a cependant pas été facile, il a souvent fallu contacter directement les collègues, les rassurer sur leur capacité à écrire sur ces sujets, qui nécessitent de se déplacer intellectuellement et convoquent une écriture et des références différentes⁶. Il était donc fondamental de pouvoir s’exprimer dans sa langue maternelle (en français, pour la plupart).

Entre temps, nous avons créé un groupe informel à la suite de l’atelier ETeRNAL de 2015, comprenant des chercheurs (Gilles Adda, Maxime Amblard, Jean-Yves Antoine, moi-même, puis Aurélie Névéol) et des industriels du TAL (Alain Couillault, Hugues de Mazancourt), avec pour but d’animer un blog sur le sujet. L’idée était à l’origine de communiquer au grand public sur le TAL, ses performances réelles et les problèmes éthiques qu’il pose. Si le blog « éthique et TAL »⁷ n’a pas été aussi vivant que nous l’aurions souhaité et n’a sans doute pas atteint le public escompté, il offre aujourd’hui à la lecture 35 articles sur des sujets variés, dont certains ont donné lieu à des publications (Névéol et al., 2017; Fort and Névéol, 2018), ce qui n’était pas prévu au départ.

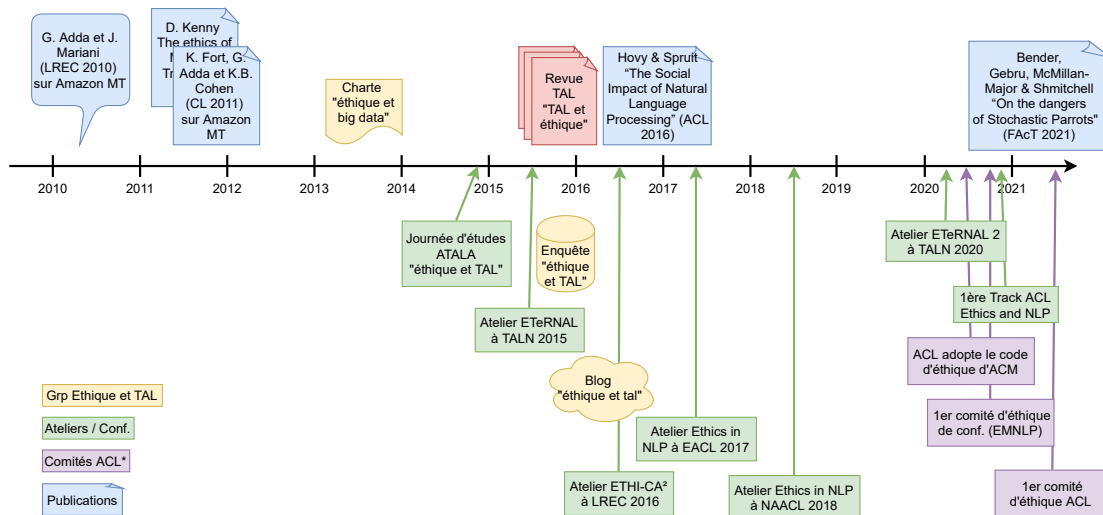


FIGURE 3.1 – Activités autour de l’éthique dans le TAL sur la dernière décennie.

J’ai également animé avec Alain Couillault (Apoliade) les discussions autour de la création de la Charte « éthique et big data » (Couillault and Fort, 2013; Couillault et al., 2014), une charte auto-déclarée qui permet d’assurer la traçabi-

6. En témoignent les remerciements de Jean-François Bonastre dans son article pour ETeRNAL 2 : « Merci à Gilles Adda et Karen Fort, pour avoir sollicité cet article (et avoir su à la fois insister et soutenir l’auteur dans ses moments de doute) ».

7. Voir : <http://www.ethique-et-tal.org>.

lité des données utilisées dans un projet, la propriété intellectuelle associée, ainsi que les moyens de leur création (producteurs de données, travailleurs) et de leur évaluation. Son impact a malheureusement été limité (seul Cap Digital l'a adoptée pour ses projets), ce qui semble être le destin de la plupart des chartes de ce type (Hagendorff, 2020). Elle a cependant permis de réunir et de faire aboutir ensemble sur un sujet complexe des acteurs variés, des industriels (GFII) aux sociétés savantes du traitement automatique des langues (ATALA) et de la parole (AFCP).

Nous avons ensuite réalisé deux enquêtes en 2015 (une nationale, l'autre internationale) sur la perception de l'éthique par les chercheurs en TAL, que nous avons présentées à LREC 2016 (Fort and Coullault, 2016). Laurence Devillers (LISN / Sorbonne Université) a organisé lors de cette même conférence le premier atelier international sur l'éthique, ETHics In Corpus Collection, Annotation & Application (ETHI-CA²).

3.1.3 D'une dynamique locale à une reconnaissance internationale

Un essor de la thématique depuis 2016

Si ces efforts ont permis de créer une dynamique locale et de favoriser des collaborations, la situation n'a guère évolué au niveau international avant l'article de Dirk Hovy et Shannon Spruit portant sur l'impact social du TAL (uniquement d'un point de vue utilisateur) publié à ACL 2016 (Hovy and Spruit, 2016), puis l'atelier *Ethics in NLP* à EACL 2017. Depuis, les questions éthiques et plus généralement les questions méta, liées au TAL, sont peu à peu devenues plus visibles et plus légitimes.

On assiste en effet depuis 2019 à une multiplication des publications sur l'éthique dans le TAL, poussée à la fois par les appels à publications et par la publicité faite sur ces sujets par certaines personnalités connues et actives dans la recherche comme sur les réseaux sociaux, notamment Emily Bender (University of Washington). Ce mouvement est parallèle à celui qui agite les médias sur l'éthique et l'intelligence artificielle et est sans nul doute lié à la révolution que vit le domaine depuis quelques années : le TAL est devenu suffisamment intéressant commercialement pour sortir des laboratoires de recherche et envahir nos vies quotidiennes, avec des conséquences immédiatement visibles pour le grand public⁸. Récemment,

8. L'agent conversationnel Tay de Microsoft, devenu raciste, sexiste et antisémite en moins de 24 h sur Twitter, a montré au grand public les limites de l'apprentissage automatique plus efficacement que n'importe quel cours introductif sur le sujet (voir : <https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?>).

3.1 Créer le chemin vers la légitimité

les capacités de GPT-3 ont attiré l'attention des grand médias et certains collègues d'autres disciplines ont commencé à s'intéresser au TAL (Floridi and Chiriatti, 2020).

Il est cependant dommage que les efforts et l'antériorité de la communauté française sur ces sujets n'aient pas été davantage reconnus à l'international. Sans doute est-ce dû à la langue des publications, qui les a rendus peu visibles. Il me semble également que nous sommes arrivés un peu trop tôt et que nous étions trop peu connus, trop peu présents sur les réseaux sociaux.

Une structuration en cours

L'année 2020 a marqué un tournant dans la reconnaissance de l'importance du sujet dans la communauté internationale, avec l'adoption par la société savante internationale du TAL, l'Association for Computational Linguistics (ACL), du code d'éthique d'ACM⁹, la première *track* spécifique d'ACL *Ethics and NLP*¹⁰ et le premier comité d'éthique de conférence à EMNLP 2020, dont j'ai été co-responsable avec Dirk Hovy. J'ai souhaité assurer la continuité du processus en acceptant la co-responsabilité du comité d'éthique suivant, celui de NAACL 2021, avec Emily Bender.

Cette évolution a culminé en 2021, avec la création par l'ACL d'un comité d'éthique pour toute l'association, dont je suis co-présidente pour cinq ans avec Min-Yen Kan (National University of Singapore) et Yulia Tsvetkov (University of Washington). Outre les trois responsables de ce comité, chaque aire géographique de l'organisation – représentées par les associations locales : AACL pour l'Asie, EACL pour l'Europe et NAACL pour l'Amérique du Nord – se voit attribué deux responsables locaux pour trois ans. Bien que la création du comité d'éthique n'ait été annoncée qu'en octobre 2021, les co-présidents ont commencé à travailler en août. La première tâche que nous nous sommes donnée a été de mettre au point un questionnaire pour évaluer la perception de l'éthique, et plus particulièrement des comités d'éthique de conférences, dans le TAL. Étant à l'origine de la première enquête sur la perception de l'éthique dans le TAL, il me semblait important de faire le point à nouveau six ans plus tard, afin de vérifier l'importance du sujet pour la communauté et de proposer des actions véritablement en lien avec ses besoins. Le travail collectif a été long, mais il a porté ses fruits et l'enquête est en cours d'analyse.

Nos priorités suivantes sont de proposer des espaces de formation à l'éthique et de réflexion collective dans la communauté, comme des tutoriels, des tables rondes

9. Voir : <https://www.acm.org/code-of-ethics>.

10. 44 papiers y ont été soumis, dont 13 acceptés, voir : <https://acl2020.org/blog/general-conference-statistics/>.

Chapitre 3. L'éthique dans et pour le TAL

ou des hackathons visant à produire du matériel pédagogique. Nous proposons également des ressources, en particulier une bibliographie participative sur l'éthique dans le TAL, déjà bien avancée ¹¹.

Nous avons également commencé à rédiger des recommandations pour les comités d'éthique de conférences du domaine. Cette dernière tâche est complexe, car nous souhaitons dépasser les simples *checklists*, certainement utiles, mais réductrices et peu à même de provoquer la réflexion. Par ailleurs, nous avons élargi la réflexion à des recommandations pour les organisateurs de conférences et les responsables de comité d'éthique, tant la manière dont on organise les lectures nous paraît au moins aussi importante que les lectures elles-mêmes. Ainsi, un comité d'éthique se doit d'être représentatif de la communauté, avec un équilibre à respecter en termes de genres, de pays, de séniorité notamment. Il doit également avoir le temps d'effectuer son travail dans de bonnes conditions et ses décisions doivent être soutenues par les organisateurs.

Ce comité d'éthique n'en est qu'à ses débuts à l'heure où j'écris ces mots, il est donc difficile d'en définir les contours exacts, mais nous le voyons davantage comme une instance d'animation de la communauté sur le sujet plutôt que comme une instance punitive.

Quoi qu'il en soit, les progrès sont bien visibles : des sujets sur lesquels il fallait argumenter pied à pied il y a une dizaine d'années, comme la rémunération des travailleurs sur **Amazon Mechanical Turk**, sont devenus presque une évidence aujourd'hui, au point que nous avons inclus sans difficulté dans la Foire Aux Questions éthiques pour NAACL 2021 l'obligation de déclarer la compensation financière prévue pour les travailleurs sur ce type de plateforme ¹².

Enfin, il est à noter que pour l'instant la communauté francophone ne s'est pas encore dotée d'un code ou d'un comité d'éthique, ce qu'elle pourrait facilement faire par le biais de sa société savante, l'Association pour le Traitement Automatique des Langues (ATALA). Cela lui permettrait non seulement de relayer les efforts réalisés à l'international mais également de participer à rendre plus visible l'antériorité de la France sur ces questions, puisque l'ATALA a participé à la rédaction de la Charte « éthique et big data » dès 2013.

3.1.4 Des oppositions peu spécifiques au TAL

Les évolutions mentionnées ici ne doivent pas masquer le fait que tous les collègues ne les voient pas d'un bon œil. Les arguments énoncés sont variés, je vais tenter de lister les plus couramment entendus.

11. Voir : <https://github.com/acl-org/ethics-reading-list/blob/main/README.md>.

12. Voir : <https://2021.naacl.org/ethics/faq/>.

La distanciation morale (*moral buffer*)

Le plus courant est que les chercheurs ne sont pas responsables des applications qui sont faites des recherches qu'ils développent. Une autre manière de le formuler est de dire que ce n'est pas aux chercheurs de se poser ces questions ¹³. Il ressort de notre enquête sur l'éthique dans le TAL réalisée en 2015 que 44,5 % des répondants ne se sentaient pas ou peu responsables des applications qui sont faites de leurs outils (Fort and Couillault, 2016). C'est ce qu'on appelle le *moral buffer* en anglais, que l'on pourrait traduire par la « distanciation morale ». Une variante de cet argument est celui qui consiste à dire que ces sujets ne relèvent pas du TAL. C'est une remarque qui m'a été faite quelques (rares) fois et que l'on retrouve indirectement dans notre enquête dans les 18,5 % de réponses négatives à la question « Pensez-vous que l'éthique doit faire partie des sujets des appels à publications des conférences du domaine ? » (Fort and Couillault, 2016). Il serait intéressant de rejouer cette enquête, afin de mesurer l'impact des évolutions récentes, en y ajoutant des questions relatives aux oppositions citées ci-dessous. Cela a d'ailleurs été ma première proposition d'action du comité d'éthique d'ACL. Quoi qu'il en soit, la responsabilité du chercheur est une question qui n'est pas nouvelle, à laquelle le procès des médecins nazis, à Nuremberg, avait apporté des réponses fermes. Le temps passe et l'oubli s'installe peu à peu ¹⁴. Il nous revient de rappeler que ni la manière dont on « fait » la science, ni les applications qu'elle permet ne sont intrinsèquement bonnes.

Le relativisme éthique

Un autre argument, que j'ai entendu porter par Pascale Fung (Hong Kong University) lors de la table ronde sur l'éthique organisée dans le cadre d'EMNLP 2020 ¹⁵, ainsi que par d'autres collègues, sur Twitter notamment, est celui du relativisme éthique : tout jugement éthique est relatif à la culture dont il est issu, il n'existe pas de valeurs universelles. Ainsi, il serait acceptable pour certains peuples de ne pas pouvoir profiter du respect à la vie privée ou à la liberté d'expression et il est impossible de décider ce qui est éthique ou non au niveau international. On peut répondre à cet argument de plusieurs manières ¹⁶, mais il me semble, très

13. La remarque m'en a été faite lors de ma présentation de l'enquête sur l'éthique à LREC 2016, j'en parle ici : <https://lstu.fr/petits-poneys-roses>.

14. Lors de mes cours sur l'éthique, je demande toujours aux étudiants (en Master ou en doctorat) s'ils ont entendu parler de ce qu'on fait ces « chercheurs en médecine » dans les camps de concentration et il y a souvent entre un quart et une moitié des présents qui n'en ont pas connaissance.

15. Voir : <https://2020.emnlp.org/schedule>.

16. Certains l'ont fait beaucoup mieux que je ne pourrais le faire, notamment Massé (2000) et surtout Macklin (1999).

Chapitre 3. L'éthique dans et pour le TAL

pragmatiquement, que le fait que l'Organisation des Nations Unies (ONU) ait fait adopter à ses 58 états membres une déclaration universelle des droits humains en 1948, est un signe fort en faveur de la reconnaissance de valeurs universelles. Le mouvement s'est par la suite amplifié puisque près de vingt ans plus tard, en 1966, deux pactes, l'un relatif aux droits civils et politiques¹⁷ et l'autre aux droits économiques, sociaux et culturels¹⁸, ont été rédigés et sont aujourd'hui ratifiés par plus de 160 pays¹⁹. Ces droits incluent entre autres l'égalité devant les tribunaux, le respect de la vie privée, le droit à l'autonomie de décision (qui inclut le consentement à une expérience médicale ou scientifique) et la liberté d'opinion. Cela ne signifie pas qu'il ne faut pas prendre en compte les éventuelles différences culturelles locales, mais qu'il faut les considérer avec un regard critique et ne pas oublier que « [...] ces choix ne sont pas faits par des individus parfaitement autonomes, les valeurs à départager étant généralement promues par des autorités sociales, morales ou politiques. » (Massé, 2000).

La crainte de la censure

Enfin, certains collègues expriment leur peur de ne plus pouvoir faire la recherche qu'ils souhaitent, d'être en quelque sorte censurés par les comités d'éthique du domaine. Cette préoccupation doit être prise en compte et discutée sérieusement. Elle ne porte cependant pas sur l'éthique elle-même mais sur les règlements dont est en train de se doter notre communauté. De fait, il est fondamental que les décisions finales concernant l'acceptation (ou non) des articles soient prises par les *chairs* des conférences et non pas par le comité d'éthique, même si celui-ci peut recommander un rejet. Par ailleurs, il est tout aussi important que les critères de décision concernant l'éthique soient clairement explicités et suffisamment publicisés. Nous concernant, le code d'éthique d'ACM est librement accessible, ainsi que, dorénavant, les FAQ éthiques des conférences²⁰. Ce n'était pas le cas pour EMNLP 2020, du fait de la mise en place du comité en urgence, au moment même de la relecture.

Cela étant, la liberté académique n'exclut en aucun cas le respect des autres, de leur liberté, de leur vie privée. Surtout, comme le dit Emily Bender dans un billet très étayé²¹ :

« [...] academic freedom comes with the responsibility of academic

17. Voir : <https://www.ohchr.org/fr/professionalinterest/pages/ccpr.aspx>.

18. Voir : <https://www.ohchr.org/fr/professionalinterest/pages/cescr.aspx>.

19. Bien entendu, cela ne signifie pas que ces droits sont respectés dans autant de pays, mais cela ne remet pas en cause l'universalité des valeurs qui les sous-tendent.

20. Voir par exemple celle de NAACL 2021 : <https://2021.naacl.org/ethics/faq/> ou d'ACL 2021 : <https://2021.aclweb.org/ethics/Ethics-FAQ/>.

21. Voir : <https://lstu.fr/academic-freedom-integrity-and-ethical-review>.

integrity. When a professional organization (such as the ACL) institutes practices of ethics review, this is an instance of the organization exercising its academic freedom to raise the standards of academic integrity within the field it represents. » ²².

Ces questions, qui tournent toutes autour de la notion de responsabilité du chercheur, ne sont en rien spécifiques au TAL, on le voit dans les références que je cite ici. Il est intéressant de se tourner vers d'autres domaines pour nourrir la réflexion, soit parce qu'ils traitent du sujet depuis longtemps, comme la philosophie, soit parce qu'ils ouvrent sur des applications différentes, comme l'informatique industrielle.

3.2 Une éthique morcelée par la spécialisation

Il est naturel en science de découper les sujets pour les traiter en profondeur. Cette manière d'appréhender les questions est d'autant plus présente en informatique que nous apprenons dès nos premiers cours d'algorithmique à « diviser pour mieux régner ». L'éthique ne fait pas exception et la communauté s'est appropriée le sujet en le découpant en morceaux plus facilement « traitables ».

Cette approche morcelée est ce que le philosophe spécialiste en éthique de l'information, Luciano Floridi, appelle la micro éthique, qu'il contraste avec la macro éthique (Floridi, 2013) et qui pose des problèmes de compartimentalisation :

« In an NLP context, an example could be a friendly and caring scientist that unwittingly abuses workers using a crowdsourcing API, because he needs gold data and has a small budget. » ²³ (Leidner and Plachouras, 2017)

Si j'essaye de prendre du recul et d'avoir une vision plus systémique des problèmes (voir Section 3.3), je participe également à l'effort micro-éthique du domaine, en produisant des ressources pour l'évaluation des outils (voir Section 3.2.1) ou en utilisant le TAL pour le TAL (*NLP4NLP*) pour examiner *comment* on publie la recherche en TAL (voir Section 3.2.2).

22. En français : « [...] la liberté académique va de pair avec la responsabilité de l'intégrité académique. Lorsqu'une organisation professionnelle (comme l'ACL) met en place des pratiques de relectures éthiques, c'est une illustration de son exercice de la liberté académique de relever le niveau des standards de l'intégrité académique dans le domaine qu'elle représente. »

23. En français : « Dans le contexte du TAL, un exemple pourrait être un scientifique amical et attentionné qui maltraite sans même s'en rendre compte les travailleurs qui travaillent pour lui via une API de microtravail, parce qu'il a besoin de données de référence et qu'il n'a qu'un petit budget. »

3.2.1 Limiter les biais

Les modèles de langue sont aujourd'hui au centre de la recherche en TAL (voir la figure 3.2). Ces mastodontes posent un certain nombre de problèmes éthiques, qui ont été mis en lumière de belle manière par Emily Bender, Timnit Gebru, Angelina McMillan-Major et Margareth Mitchell dans un article qui a fait sensation (Bender et al., 2021). Un de ces problèmes est celui des biais stéréotypés dont ces modèles se font non seulement l'écho, mais qu'ils amplifient (Zhao et al., 2017). Ce sujet éthique est celui qui a le mieux « pris » dans le domaine. La majeure partie de l'article de Hovy and Spruit (2016) porte d'ailleurs sur des sujets liés aux biais : biais d'exclusion, de sur-généralisation, de surexposition ou de sous-exposition (ce dernier étant lié à la diversité linguistique). La série d'ateliers « Fairness, Accountability, and Transparency in Machine Learning (FATML) », qui a débuté en 2014 (FAT, 2014) semble être la première conférence organisée sur le sujet.

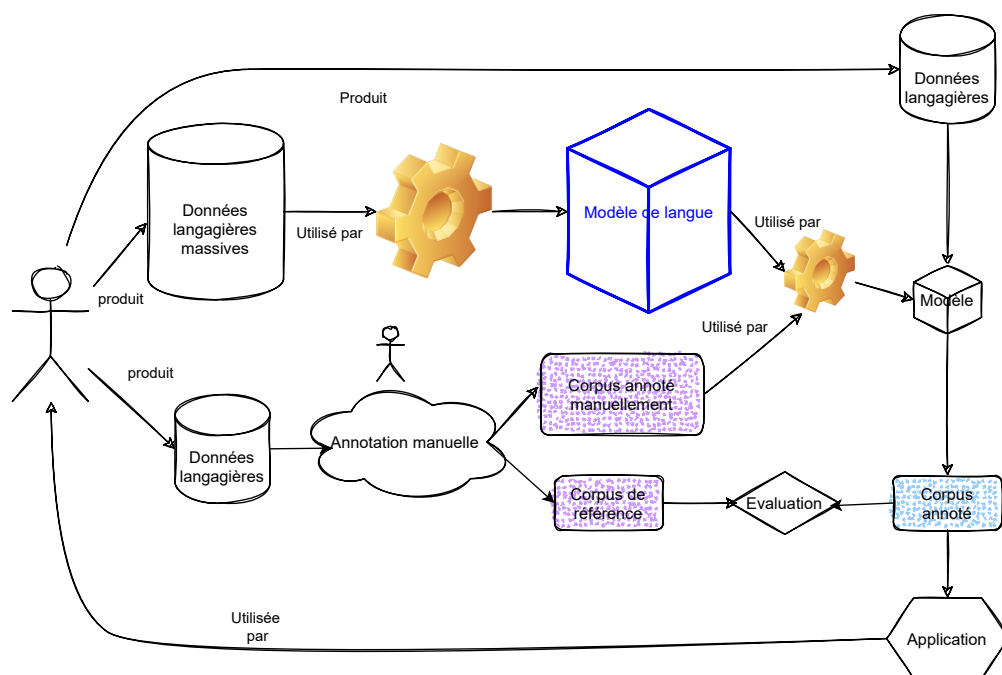


FIGURE 3.2 – Les modèles de langue au centre du TAL actuel.

L'agitation autour de ce sujet a d'ailleurs donné lieu à une très riche méta-étude, celle de Blodgett et al. (2020), pour laquelle les auteurs ont analysé pas moins de 146 papiers portant sur le sujet depuis 2016.

Dirk Hovy a publié en 2021 un autre article centré sur le sujet (Hovy and Prabhumoye, 2021) qui recense cinq sources potentielles de biais dans le TAL

actuel (voir Figure 3.3) : l'annotation, la sélection des données, les représentations fournies en entrée, les modèles et la conception de l'expérience elle-même.

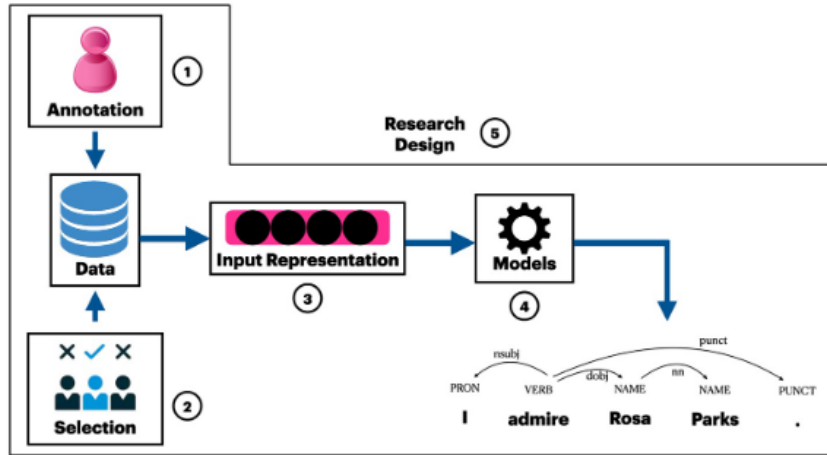


FIGURE 1 Schematic of the five bias sources in the general natural language processing pipeline

FIGURE 3.3 – Les cinq sources de biais dans le TAL actuel selon (Hovy and Prabhumoye, 2021).

Ce travail est salutaire, car il met en avant le fait que les problèmes sont multifactoriels et ne peuvent pas être réglés en ne s'attaquant qu'à une source de biais.

Dans tous les cas, pour parvenir à limiter ces biais encore faut-il pouvoir les évaluer avec précision. Nous avons besoin pour cela de ressources langagières adaptées. Des jeux de tests ont été créés pour l'anglais, notamment CrowS-Pairs (Nangia et al., 2020), mais il n'existait aucune ressource pour le français jusqu'en 2021.

J'ai travaillé cette année-là sur le sujet avec Aurélie Névél et Yoann Dupont (ATER à Sorbonne Université) dans le cadre du mémoire de M1 de Julien Bezançon. Nous avons créé un corpus biaisé/anti-biaisé (par exemple : « Les femmes sont incapables de faire des sciences » / « Les hommes sont incapables de faire des sciences ») pour le français à la manière de CrowS-Pairs. Nous avons pour cela mis en place une méthodologie applicable pour de nombreuses langues, qui consiste d'une part à traduire le corpus anglais (lorsque cela est possible et pertinent) et d'autre part à utiliser une plateforme de sciences participatives²⁴ pour collecter de nouvelles phrases adaptées à la culture considérée. Grâce aux participants, nous

24. Language Arc : <https://languagearc.com/>).

avons collecté 212 phrases représentant des biais stéréotypés de 10 types différents : ethnie, genre, orientation sexuelle, religion, âge, nationalité, handicap, statut socio-économique, apparence physique et autre (principalement orientation politique). Nous avons non seulement traduit, mais corrigé le corpus CrowS-Pairs et nous lui avons ajouté les 212 paires de phrases créées par la communauté francophone, pour un total de 1 677 paires.

Nous avons ensuite utilisé ce corpus de test pour évaluer les modèles disponibles pour le français, en particulier les plus performants, CamemBERT (Martin et al., 2020) et FlauBERT (Le et al., 2020), ce dernier générant significativement moins de phrases biaisées. Ce travail a fait l'objet d'un article publié à ACL 2022 (Névél et al., 2022), la conférence A* du TAL.

Cette première expérience nous a permis de valider notre méthodologie. Notre but est maintenant de l'élargir à de nouvelles langues pour lesquelles des modèles de langue masqués de type BERT existent. À plus long terme, il nous faudra réfléchir à des moyens efficaces pour tester d'autres types de modèles de langues, notamment les modèles génératifs de type GPT. Il faudra aussi valoriser l'évaluation éthique des modèles dans la communauté, afin de la généraliser. Mon action en tant que coprésidente du comité d'éthique d'ACL vise justement à pousser ce type de bonnes pratiques dans la communauté.

3.2.2 Améliorer la diversité linguistique

Ces dernières années ont vu la question du respect de la diversité linguistique et des droits des populations parlant des langues peu ou pas du tout dotées (d'outils, de ressources langagières) être mise sur le devant de la scène.

Joshi et al. (2020) ont ainsi réalisé une analyse très approfondie de la présence des langues dans les principales conférences de TAL. Leur travail permet de se rendre compte que d'importants progrès ont été réalisés depuis les années 2000, mais que LREC et les ateliers (*workshops*) restent de loin les plus inclusifs en termes de langues. Or, LREC n'est pas très bien classée comme conférence (CORE C) et les ateliers ne sont en général pas considérés à leur juste valeur. Par conséquent, même si des progrès sont visibles, le traitement de langues diverses n'est toujours pas considéré comme central dans le TAL, malgré les affirmations selon lesquelles les modèles actuels seraient « agnostiques » concernant les langues traitées. De ce point de vue, le plus porteur d'espoir selon moi est la création d'une communauté active autour des langues peu dotées, à travers le groupe SIGUL²⁵ (*Special Interest Group : Under-resourced Languages*), qui organise les ateliers SLTU (*Spoken Language Technologies for Under-resourced languages*) et CCURL (*Collaboration and Computing for Under-Resourced Languages*). Ce travail est no-

25. Voir : <http://www.elra.info/en/sig/sigul/>.

3.2 Une éthique morcelée par la spécialisation

tamment porté par des personnalités comme Claudia Soria (CNR-ILC) et Laurent Besacier (GETALP-LIG, maintenant NAVER LABS).

Steven Bird (Charles Darwin University) est allé plus loin en publiant à COLING 2020 un article intitulé « Decolonising speech and language technology » (Bird, 2020), dans lequel il propose des solutions pour « décoloniser » notre approche des langues indigènes, en co-construisant les technologies avec les communautés. Alice Millour est allée dans ce sens dans sa thèse, que j’ai encadrée, grâce à ses plateformes de sciences participatives (Millour and Fort, 2018c,b). En effet, la nécessité de trouver des participants, de les motiver, nous a amenées à mieux « écouter » leurs envies, leurs blocages (Millour and Fort, 2018a; Millour, 2019). Nous étions encore loin de la co-construction, cela dit, mais pour cela il aurait fallu être intégrées dans la communauté de locuteurs, et ce, dès le début de la recherche. Le jeu Katana et Grand Guru que nous avons créé dans le cadre de l’action COST enet-Collect était un effort en ce sens puisqu’il est conçu pour favoriser les interactions intergénérationnelles dans la langue parlée par les aînés (Millour et al., 2019), un besoin exprimé par les locuteurs.

Au-delà de ces efforts, le domaine du TAL reste fortement marqué par une focalisation sur quelques langues, en particulier l’anglais. Cette focalisation est telle que l’anglais est devenu, en quelque sorte la « langue par défaut », celle que l’on n’a même pas besoin de citer lorsque l’on travaille dessus. Cette construction a un impact négatif sur la diversité linguistique, puisqu’elle tend à faire croire que seul l’anglais est important ou que si l’on obtient de bons résultats sur cette langue alors cela va fonctionner avec toutes les langues. Cet état de fait a été dénoncé avec force par E. Bender (Université de Washington) d’abord dans un article en 2011 (Bender, 2011), puis dans un billet de blog en 2019²⁶, dans lequel elle énonce la règle dite « de Bender » :

« Always name the language(s) you’re working on. »²⁷

J’ai mené récemment un travail, en collaboration avec Yves Lepage (Université de Waseda, Japon), Gaël Lejeune (Sorbonne Université) et une étudiante de L3 puis M1, Fanny Duce, visant à quantifier l’application de cette règle dans les articles des conférences en TAL. Nous avons pour cela créé un corpus de près de 15 000 articles de recherche issus des conférences ACL (7 262 articles), LREC (6 669 articles) et TALN (1 172 articles). Nous avons annoté manuellement plus de 1 500 articles selon que les auteurs citent ou non la ou les langues sur lesquels ils travaillent et si cette information est déductible du fait des ressources langagières citées. Nous avons ensuite utilisé cette référence pour entraîner des classifieurs, que nous avons appliqués sur tout le sous-corpus (en français pour TALN, en anglais pour les autres conférences).

26. Voir : <https://lstu.fr/bender-rule>.

27. En français : « Il faut toujours spécifier la ou les langues étudiées. ».

Chapitre 3. L'éthique dans et pour le TAL

Il ressort de notre annotation manuelle que les langues utilisées et non citées sont, sans surprise, l'anglais pour ACL et LREC, et l'anglais et le français pour TALN.

Les résultats obtenus par les classifieurs en diachronie sont présentés dans la figure 3.4. Il est à noter que LREC a lieu tous les deux ans et que ses actes ne sont disponibles que depuis 2000. Par ailleurs, nous n'avons utilisé pour TALN que les articles de l'archive TALN en français, au format texte (Boudin, 2013), qui ne sont disponibles que depuis 2000.

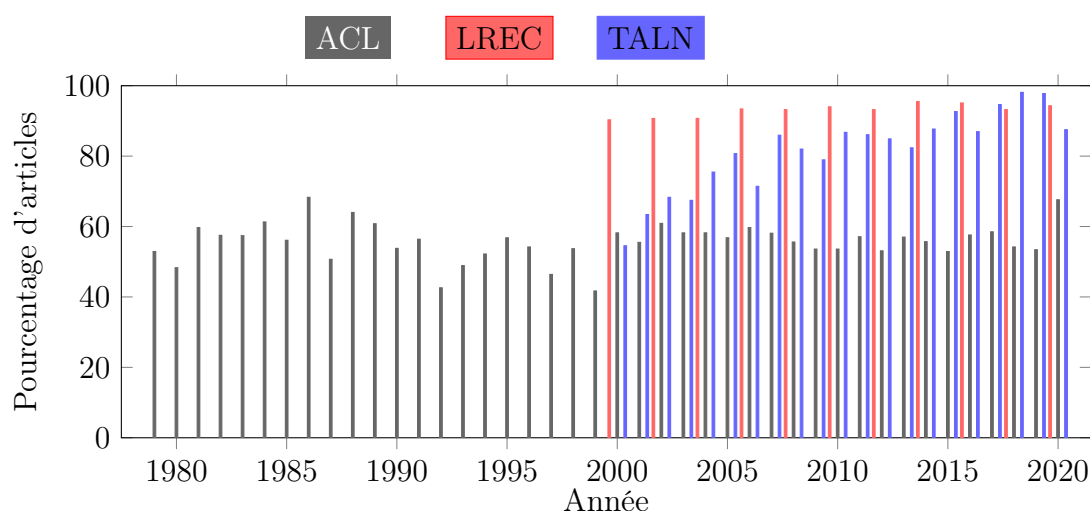


FIGURE 3.4 – Proportion d'articles appliquant la règle de Bender pour chaque édition des trois conférences (ACL en noir, LREC en rouge et TALN en bleu).

Comme le montre la figure 3.4, au moins 30 % des articles d'ACL (la conférence A* du domaine) n'appliquent pas la règle de Bender, alors que plus de 80 % des articles de LREC la respectent (LREC étant une conférence spécialisée dans les ressources langagières, cela semble logique). Nous observons cependant un sursaut à ACL 2020, peut-être dû au billet de blog d'Emily Bender (il faudra vérifier la tendance dans les années à venir), la situation peut donc évoluer rapidement.

En ce qui concerne TALN, la proportion d'articles respectant la règle de Bender a pratiquement toujours été supérieure à celle d'ACL (sauf en 2000) et dépasse même celle de LREC ces dernières années. Ces relativement bons résultats sont peut-être à mettre au crédit de l'écriture en français, qui n'est pas la lingua franca de la recherche, ce qui nous pousserait à définir la ou les langues sur lesquels nous travaillons de manière explicite.

Ces travaux sont détaillés dans deux articles, l'un à LREC 2022, portant sur les conférences ACL et LREC, et l'autre à TALN 2022, ciblant ACL et TALN. La tendance est suffisamment nette pour qu'on puisse conclure sur une prévalence

toujours importante dans la principale conférence du domaine d’articles ne citant pas la langue sur laquelle porte les travaux. Je pense néanmoins qu’il faut rester optimiste et que la sensibilisation des relecteurs sur ce sujet et l’influence des réseaux sociaux peut faire évoluer la situation rapidement.

3.2.3 Documenter les données

La nécessité de documenter les ressources langagières est reconnue depuis longtemps en linguistique, en particulier en linguistique de terrain (Bouquiaux and Thomas, 1971) et en linguistique de corpus (Leech, 1991; Wynne, 2005). La langue orale n’est pas en reste, en témoigne le livre sur les bonnes pratiques pour les corpus oraux publié en 2006 (Baude et al., 2006). La Text Encoding Initiative (TEI) a amplifié le phénomène en promouvant la documentation des sources et des annotations dans ses *headers* (Burnard, 2005). Bender and Friedman (2018) approfondissent ces bonnes pratiques en proposant une documentation détaillée des données (*data statements*) visant à réduire les biais dans les systèmes de TAL produits à partir de celles-ci. Mitchell et al. (2019) poursuivent le même but et proposent des cartes d’identité précises pour documenter les modèles d’apprentissage dont deux parties concernent les données : « *Evaluation data* » (données pour l’évaluation) et « *Training data* » (données pour l’entraînement).

Si ces efforts sont à saluer, ils ne couvrent cependant qu’une partie de la question. En effet, aucune de ces bonnes pratiques ne mentionne l’importance de la traçabilité des ressources : est-ce que la ressource est primaire ? Si elle ne l’est pas, est-ce que la ressource primaire est documentée et, si oui, où ? quels sont les droits associés à cette ressource ? Il en va de même concernant le statut des participants à la ressource : sont-ce des étudiants, des travailleurs, des collègues ? ont-ils été rémunérés ? comment ? Enfin, il est encore rare en linguistique de donner un accord inter-annotateurs, la notion de qualité est donc peu abordée.

C’est en réaction à ces manques que nous avons créé avec Alain Couillault (à l’époque délégué général au Groupement Français de l’Industrie de l’Information, GFII) la charte « Éthique et big data » (Couillault and Fort, 2013; Couillault et al., 2014), en collaboration avec l’Association pour le traitement automatique des langues (ATALA), dont j’étais la représentante, l’Association française de la communication parlée (AFCP), représentée par Gilles Adda, et Cap Digital, représenté par Christelle Ayache. Cette charte comprend quatre parties²⁸ :

1. description des données,
2. traçabilité,
3. propriété intellectuelle,

28. Voir : <https://lstu.fr/charte-ethique-et-big-data>.

4. réglementations spécifiques

La partie traçabilité est la plus fournie, puisqu'elle comprend six sous-parties, concernant non seulement l'origine des données, les auteurs et leur recrutement, mais également les processus de fabrication et de validation des données. Cette charte est auto-déclarée. Notre but était de la faire prendre en compte par les organismes de financement tels que l'ANR et de la rendre obligatoire à chaque dépôt de projet. Malheureusement, seul Cap Digital l'a adoptée.

3.2.4 Limiter l'empreinte carbone

La question de l'impact environnemental de nos outils est apparue très récemment, en particulier avec l'article de [Strubell et al. \(2019\)](#). Un premier atelier sur le sujet, SustaiNLP ([Moosavi et al., 2020](#)) a été organisé dans le cadre d'EMNLP 2020 et a été reconduit l'année suivante ([Moosavi et al., 2021](#)). La richesse des actes de ces ateliers est la preuve que la communauté du TAL s'intéresse au sujet. Il reste cependant difficile à faire prendre en compte concrètement : en tant que co-présidentes du comité d'éthique de NAACL 2021, nous avons proposé, Emily Bender et moi-même, de demander aux auteurs d'indiquer l'empreinte carbone de leur recherche, mais cela nous a été refusé et nous n'avons pu qu'inciter à le faire²⁹. Il est vrai que ce type de mesure n'est pas si facile à réaliser et que nous manquions d'un état de l'art des outils adéquats, qui a été produit entre temps dans [Bannour et al. \(2021\)](#). Il y a fort à parier que présenter ces informations devienne une évidence dans les années à venir.

Je ne vais pas m'étendre davantage sur ce sujet, sur lequel je ne travaille pas directement, mais je souhaite ici ajouter un point qui me semble important et qui a été mis en lumière par Emily Bender lors d'un exposé au Turing Institute en juillet 2021. Certains collègues soutiennent que les très grands modèles de langues permettent justement de moins consommer d'énergie puisqu'on peut les utiliser tels quels en les « affinant » (*fine tuning*) en fonction des applications. C'est effectivement le cas. Néanmoins, la succession rapprochée de productions d'énormes modèles de langues que montre la figure 3.5³⁰ illustre l'échec de cette stratégie. En effet, non seulement on en compte une douzaine sur environ deux ans, mais ces modèles concernent surtout l'anglais³¹. Ils sont produits pour la très grande majorité par des entreprises du TAL, qui n'hésitent pas à consommer énormément

29. Voir : <https://2021.naacl.org/ethics/faq/>.

30. Cette figure est extraite (avec son accord) de l'exposé d'E. Bender au Turing Institute en juillet 2021 sur ([Bender et al., 2021](#)) (figure non présente dans l'article), voir : <https://faculty.washington.edu/ebender/papers/Bender-Turing-Institute-July-2021.pdf>.

31. Pour le français, on en compte déjà deux : CamemBERT ([Martin et al., 2020](#)) et FlauBERT ([Le et al., 2020](#)), produits quasiment en parallèle.

3.3 Pour une vision plus systémique de l'éthique dans le TAL

d'énergie pour entraîner des modèles toujours plus énormes, afin d'en faire la publicité et entretenir une image compétitive. C'est sans aucun doute un effet néfaste de la compétition féroce que se livrent ces entreprises. Je traiterai d'autres effets dans la section 3.3.3.

On peut cependant espérer que la tendance se calme dans les années à venir pour les langues disposant déjà de modèles.

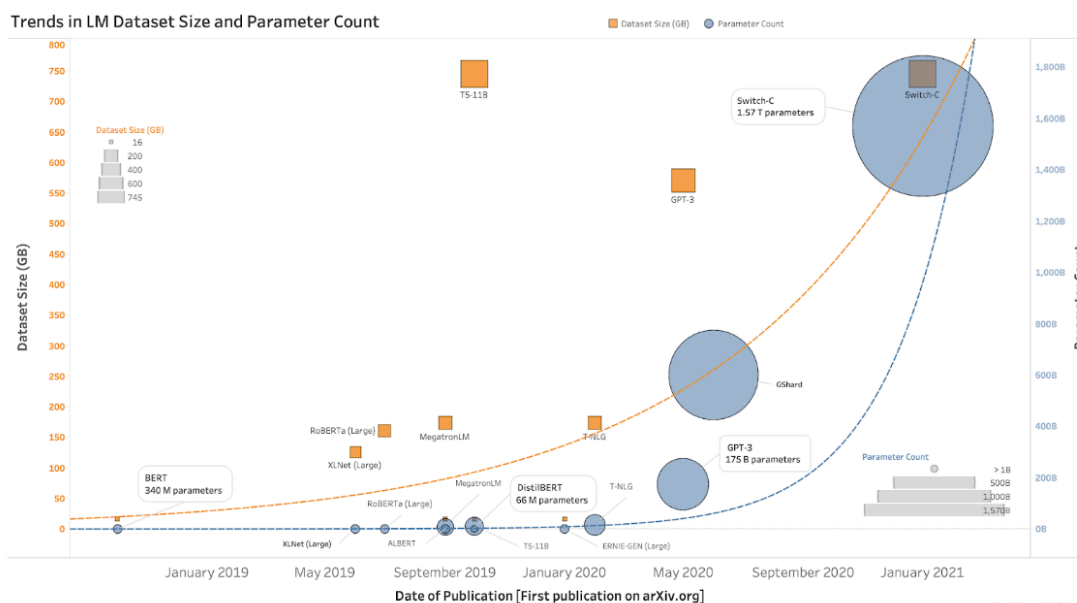


FIGURE 3.5 – Évolution de la production des modèles de langues entraînés ces dernières années en taille du jeu de données d'entraînement et en nombre de paramètres, schéma issu d'une présentation de l'article (Bender et al., 2021).

Si les thématiques que j'ai présentées rapidement ici sont actuellement très vivaces en TAL, elles sont loin de couvrir tout le spectre des questions éthiques qui devraient se poser dans le domaine. Pour identifier les angles morts (*blind spots*), il faut prendre du recul et rechercher un point de vue plus élevé.

3.3 Pour une vision plus systémique de l'éthique dans le TAL

Pour sortir du morcellement, il faut pouvoir mobiliser une grille d'analyse. Il m'a donc fallu (re)plonger dans les grandes visions philosophiques de l'éthique. J'ai pour ce faire suivi en 2018 le MOOC Coursera de l'Université de Genève : *Le*

*Bien, le Juste, l'Utile. Introduction aux éthiques philosophiques*³² Ce MOOC est très bien conçu, sur la forme comme sur le fond, et m'a permis d'élargir ma vision des questions éthiques dans le domaine du TAL et de mieux comprendre ce qui les relie et comment. J'ai complété cette formation par le MOOC edX de l'Université du Michigan sur l'éthique de la science des données (*Data Science Ethics*), qui est uniquement conséquentialiste. J'ai pu grâce à lui vérifier que je disposais désormais des outils pour analyser les situations d'un point de vue éthique. Bien entendu, cette appropriation d'un nouveau domaine ne s'arrête pas à ces formations. Je me nourris de nombreuses lectures et interactions, notamment avec mon post-doctorant philosophe, Marc Anderson.

3.3.1 Les éthiques philosophiques

Il existe trois principaux courants éthiques en philosophie, que je présente ici rapidement.

Le premier, historiquement, est l'éthique des vertus, qui a été développée par Aristote dans l'Éthique à Nicomaque autour de 350 av. JC ([Aristote, 350](#)). Aristote fonde sa réflexion sur l'eudémonisme, ou le fait que le bonheur souverain est le but de la vie humaine. Or, ce bonheur dépend de la *pratique* des vertus, en particulier de la plus importante selon Aristote, la prudence (tempérance), c'est-à-dire le juste milieu. L'être humain doit développer sa raison et ses vertus par la pratique et dans un cadre social, jusqu'à en devenir virtuose.

On retrouve ce perfectionnisme chez Emmanuel Kant dans les Critique de la raison pure ([Kant, 1781](#)) et Critique de la raison pratique ([Kant, 1788](#)) qui fondent l'éthique déontologique. Pour Kant, en effet, l'être humain doit avoir pour but de devenir meilleur et doit pour cela se soumettre à un principe moral qui intervient a priori et qui est absolu. C'est la soumission à ce devoir, à cette loi interne, qui nous élève. Pour être réellement libre l'être humain doit raisonner (en pratique) et agir en cohérence, sans être l'esclave de ses passions. Il doit avoir le souci des autres et penser en termes de la « bonne » action. L'éthique déontologique reste d'actualité, avec des déclinaisons contemporaines notamment autour de Jürgen Habermas (éthique du discours) et Hans Jonas (éthique de la responsabilité, principe de précaution).

Enfin, le courant éthique le plus convoqué actuellement est sans aucun doute l'éthique utilitariste de Jeremy Bentham ([Bentham, 1780](#)) et John Stuart Mill ([Mill, 1859](#)) et son héritière, l'éthique conséquentialiste. Ce courant se veut fondé sur une méthode scientifique dans laquelle des points sont associés à chaque effet (positif et négatif) produit par une décision à prendre, qui devra donc dépendre du total obtenu. Il est à noter que dans cette théorie, personne n'est plus important qu'un

32. Voir : <https://www.coursera.org/learn/ethique>.

3.3 Pour une vision plus systémique de l'éthique dans le TAL

autre, chacun comptant pour un (égalitarisme). Elle va évoluer dans le temps : si le but pour Bentham est de maximiser le plaisir, Mill revient lui à la notion de bonheur et y ajoute la vertu. À la différence des courants précédemment décrits, l'utilitarisme n'est pas perfectionniste et ne prend en compte que les conséquences des actions réalisées.

3.3.2 Des grilles d'analyse conséquentialistes pour le TAL

L'éthique conséquentialiste sous-tend plus ou moins explicitement la plupart des recherches en éthique dans le TAL. En effet, biais, impacts environnementaux, *dual use*, manque de diversité linguistique, sont autant de conséquences néfastes sur la société que nous devons prendre en compte.

Il existe cependant très peu d'analyses conséquentialistes revendiquées comme telles, et encore moins qui proposent une vision cohérente large. La grille d'analyse proposée dans Lefeuvre et al. (2015), puis approfondie et instanciée sur les commandes vocales dans Lefeuvre-Halftermeyer et al. (2016), est de ce fait particulièrement intéressante. Elle se présente sous la forme d'une typologie contenant jusqu'à cinq niveaux de hiérarchie qui prend en compte les impacts (positifs et négatifs) sur l'individu et la société. Il ne fait pour moi aucun doute que cette étude serait très largement citée³³ si elle était rédigée ou traduite en anglais.

Récemment, un article portant une analyse de type conséquentialiste sur les problèmes que posent les modèles de langues (Bender et al., 2021) a provoqué le licenciement de deux de ses autrices, Timnit Gebru et Margareth Mitchell, par leur employeur, Google³⁴. Cette étude reste cependant concentrée sur une partie, certes centrale, du domaine et ne couvre pas la totalité du spectre détaillé dans Lefeuvre et al. (2015) (par exemple, les problèmes psychiques ne sont pas abordés). Néanmoins, c'est à ma connaissance la première publication qui met au jour le « racisme environnemental » lié aux gigantesques modèles de langues : ceux à qui ils profitent le plus (les plus privilégiés) ne sont pas ceux qui vont subir le plus les conséquences des dérèglements climatiques auxquels ils participent (les populations marginalisées).

3.3.3 Vers une grille d'analyse systémique pour le TAL

Ces recherches sont extrêmement stimulantes et permettent d'avoir une vision plus large des questions d'éthique dans le TAL. Cependant, elles se concentrent

33. À l'heure actuelle, les deux articles sont cités trois et six fois respectivement selon Google Scholar.

34. Voir : <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>.

sur les conséquences de nos activités et ne permettent pas de rendre compte de la *manière* dont on fait ces recherches et des problèmes que cela pose.

Pour cela, il faut « penser par le milieu » comme le dit Isabelle Stengers ([Stengers, 2019](#)), ce qui implique de poser le contexte, de déplier ce que nous faisons sans même y penser. Nous avons proposé, avec Maxime Amblard (Maître de conférences à l'Université de Lorraine), un début d'analyse systémique ([Fort and Amblard, 2018](#)) de l'éthique dans le TAL. Nous avons pour cela commencé par décrire l'environnement de production de la recherche, les acteurs, puis les données (voir Figure 3.6). Le haut de la figure représente la production de nos outils, le bas l'utilisation de ceux-ci et la partie gauche le financement. Les flèches bleues en pointillés représentent l'inversion possible des fonctions (les chercheurs/ingénieurs peuvent également être des producteurs de données), les flèches vertes en pointillés indiquent les biais induits par l'intervention des acteurs. Cette manière de décrire le domaine permet de mettre au jour des « angles morts » concernant l'éthique.

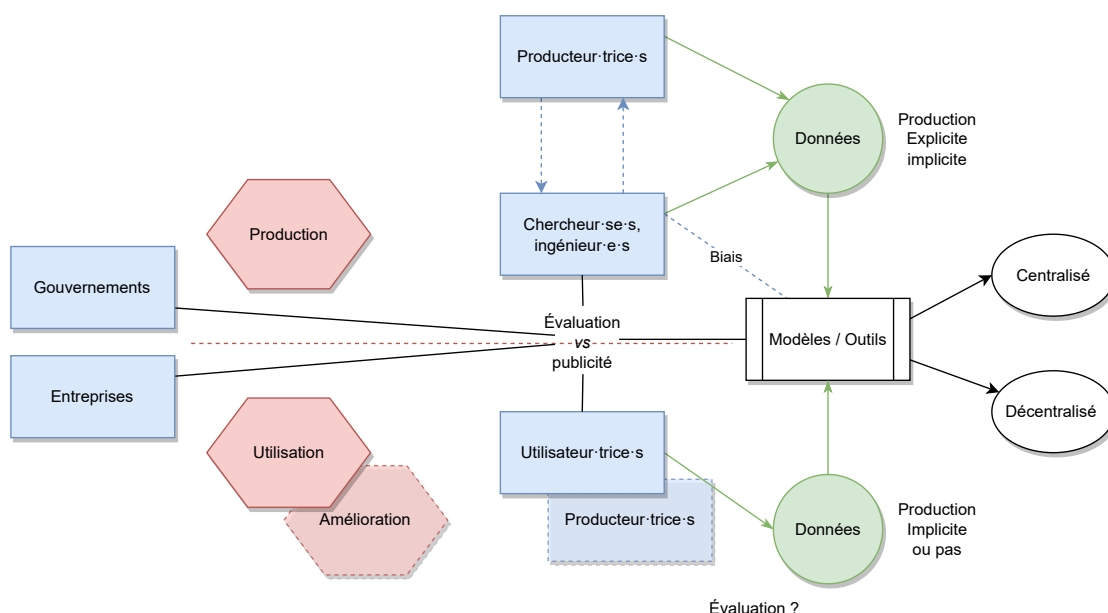


FIGURE 3.6 – Environnement de production de la recherche (rose et blanc), acteurs (bleu), données (vert).

Des utilisateurs producteurs

Nous nous sommes ainsi très rapidement rendus compte que la frontière entre les producteurs de données et les utilisateurs était de plus en plus fluide. Ce brouillage est lié à la décentralisation de nombreux outils, qui apprennent directement, de manière décentralisée, auprès des utilisateurs finaux (ce que j'appelle

3.3 Pour une vision plus systémique de l'éthique dans le TAL

le « machine learning in the wild »). C'est le cas par exemple d'un assistant virtuel qui continue à apprendre de ses échanges avec le monde qui l'entoure, comme Tay de Microsoft. Cette disparition de la frontière entre production et utilisation pose de nombreux éthiques, notamment celle du consentement des utilisateurs qui deviennent, de fait, des producteurs de données. Cela modifie également la temporalité de la production, qui échappe à tout contrôle (comme on l'a vu dans le cas de Tay), ce qui pose à son tour la question de la responsabilité : qui fait quoi ? qui est responsable de quoi ?

L'influence des « Big Tech »

Ce type de représentation permet également de visualiser un acteur clé du domaine, les grandes entreprises du TAL, dont la présence, voire l'omniprésence, a un impact important, en particulier sur l'évaluation des outils, comme on a pu le constater lors du licenciement de Timnit Gebru et Margareth Mitchell. Cette omniprésence participe sans doute également de l'accélération du domaine, car une entreprise a des impératifs en termes de publicité et se doit de montrer rapidement (et régulièrement) ce qu'elle est capable de produire, par exemple de nouveaux modèles de langue, toujours plus gros et plus performants (voir Figure 3.5). Les gouvernements ne sont certainement pas en reste, mais leur temporalité n'est pas tout à fait la même : les résultats peuvent attendre quelques années (trois à quatre pour un projet ANR, par exemple), alors que les entreprises, même grosses, n'ont que quelques mois à un an pour produire des résultats promouvables.

Nous avons choisi de représenter séparément les gouvernements (dont le monde académique est directement dépendant) et les entreprises, en tant que financeurs, mais les intrications entre les deux se multiplient à d'autres niveaux. C'est le cas en particulier au niveau de l'activité de recherche, avec la multiplication des doubles affiliations entreprise/Université qui touche le domaine du TAL, en particulier à l'international. Ce bouleversement n'est à ma connaissance pris en compte dans aucune analyse éthique du TAL à l'heure actuelle, mais il a récemment fait l'objet d'un article provocateur dans une conférence sur l'IA, l'éthique et la société ([Abdalla and Abdalla, 2021](#)). Les auteurs analysent l'impact des « Big Tech » (Google, Amazon, Facebook, Microsoft, Apple, Nvidia, Intel, IBM, Huawei, Samsung, Uber, Alibaba, Element AI, OpenAI) à travers les financements reçus par les chercheurs en IA et plus particulièrement en éthique de l'IA. La force de cet article réside dans la comparaison qu'il propose avec les manœuvres des entreprises du tabac (« Big Tobacco »), qui ont influencé la recherche sur les méfaits du tabac en finançant des universités et des chercheurs et en noyant la recherche sous leurs propres articles. Au vu de la liste des entreprises concernées, il me semble évident qu'il faut appliquer la même méthodologie sur le sous-domaine du TAL, en l'élargissant aux autres types d'influences, notamment la présence de chercheurs des « Big Tech »

Chapitre 3. L'éthique dans et pour le TAL

dans les organisations de conférence et les instances du domaine. Ce travail est en cours avec des collègues, je le présente brièvement dans les perspectives (voir Section 4.3.1).

Consentement et traçabilité : fondus dans la massification ?

La figure 3.7 permet enfin de mieux comprendre à quel point les questions du consentement et de la traçabilité des données restent complexes.

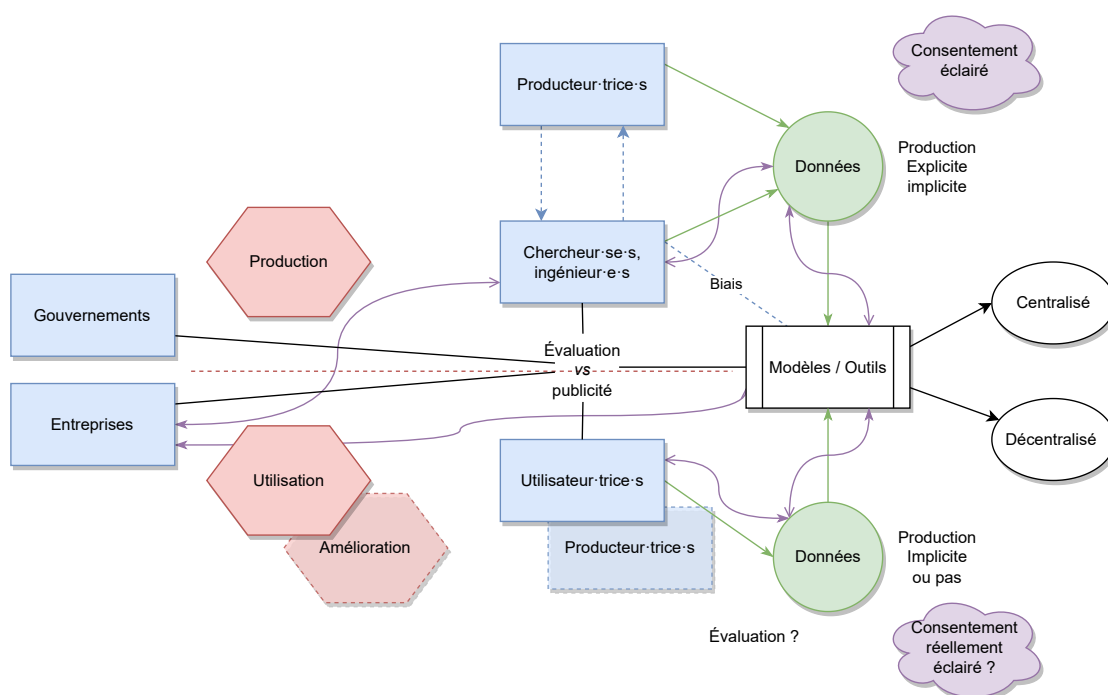


FIGURE 3.7 – Les flèches violettes montrent les flux de données et illustrent la difficulté du consentement et de la traçabilité.

En effet, le brouillage des frontières entre utilisateurs et producteurs de données ne s'est pas accompagné d'une évolution du recueil du consentement, qui n'est pas assez dynamique pour prendre en compte ces évolutions.

Par ailleurs, la masse de données nécessaire pour entraîner un modèle de langue peut servir d'excuse pour « oublier » les droits de producteurs de données textuelles sur leurs contenus. Ainsi, les corpus de CommonCrawl³⁵ ont notamment été utilisés pour entraîner les modèles de langue du français CamemBERT (Martin et al., 2020) et, dans une moindre mesure, FlauBERT (Le et al., 2020). Or, ces corpus

35. Voir : <https://commoncrawl.org/>.

3.3 Pour une vision plus systémique de l'éthique dans le TAL

ne sont disponibles que selon le principe du *fair use*, une notion du droit anglo-saxon, qui n'existe pas en France. Ils contiennent des données extraites du Web de manière indiscriminée, donc non respectueuse des licences associées. Le règlement général sur la protection des données (RGPD) ayant inversé la charge de la preuve, il faudrait qu'un internaute prouve que ses données ont été utilisées sans son accord pour obtenir un retrait de celles-ci du corpus en Europe. Si la légalité de la question est donc pour l'instant assurée, il n'en va pas de même pour l'éthique, le consentement éclairé étant partie intégrante de tous les textes concernant l'éthique de la recherche, depuis le code de Nuremberg³⁶, le rapport Belmont (Bel, 1979) jusqu'au rapport Menlo (Dittrich and Kenneally, 2012).

En ce qui concerne la traçabilité des données, celle-ci a été elle-aussi largement mise à mal par la massification, qui nécessite des efforts considérables à qui voudrait l'assurer. Le projet BigScience a semble-t-il produit de gros efforts en ce sens. Il faudra en mesurer les résultats lorsqu'ils seront disponibles.

Un travail sur et dans le temps

Ce travail est encore en cours, mais il fournit une base de réflexion originale. Nous creusons actuellement une des dimensions les plus importantes à nos yeux, celle de la manière dont nous présentons nos résultats de recherche (autrement dit, l'évaluation *vs* la publicité). Nous travaillons sur ce sujet avec Fanny Ducel, dans le cadre de son mémoire de M1, dont je détaille les pistes dans les perspectives, dans la section 4.3.1. Dans la même veine, j'ai mené cette année avec Alice Millour, Yoann Dupont et une étudiante de M1, Alexane Jouglar, une recherche sur l'évaluation des outils d'extraction d'entités nommées, qui a donné lieu à un article à TALN 2022 (Millour et al., 2022). Nous y confirmons que les performances affichées par les outils ne correspondent pas à la réalité, en particulier lorsque l'on s'écarte du domaine du corpus d'entraînement. Surtout, nous proposons un corpus en français équilibré en genres et de nouvelles visualisations qui permettent d'affiner l'évaluation des outils.

Enfin, nous avons commencé à mener une analyse temporelle du domaine, qui permet de montrer la dynamique à l'œuvre dans les dernière décennies. Nous sommes en effet passés très rapidement (en une vingtaine d'années) des systèmes experts, où seuls ceux-ci créent des outils, à l'apprentissage machine (statistique, puis neuronal), où l'expertise n'est plus à la base de la construction des outils. Ce passage à l'échelle s'est accompagné d'une accélération extrême du domaine, liée à plusieurs facteurs, dont une meilleure robustesse des outils, qui a permis d'en faire des produits plus facilement vendables, mais également un fonctionnement de la recherche orienté sur la publication immédiate (« *publish or perish* »).

36. Voir : https://media.tghn.org/medialibrary/2011/04/BMJ_No_7070_Volume_313_The_Nuremberg_Code.pdf.

Chapitre 3. L'éthique dans et pour le TAL

Le temps est le critère fondamental de l'éthique. En effet, sans temps suffisant pour penser ce que l'on fait, il ne peut y avoir d'éthique. Je ne pouvais par conséquent pas finir ce chapitre sans citer le philosophe Hartmut Rosa, spécialiste du temps :

« Les sujets modernes peuvent [...] être décrits comme n'étant restreints qu'à minima par des règles et des sanctions éthiques, et par conséquent comme étant « libres », alors qu'ils sont régentés, dominés et réprimés par un régime-temps en grande partie invisible, dépolitisé, indiscuté, sous-théorisé et inarticulé. » (Rosa, 2012)

Autrement dit, la réduction du temps est devenue notre maître et il nous faut lutter pour conserver des espaces de réflexion éthique. C'est ce que j'ai participé à amorcer ces dernières années à travers la journée d'étude ATALA « éthique et TAL », le blog éthique et TAL ³⁷, le numéro spécial de la revue TAL et les ateliers ETeRNAL. C'est ce que je continue à faire aujourd'hui, au niveau international dans le cadre du comité d'éthique d'ACL et au niveau local, dans le cadre du comité d'éthique de la recherche de Sorbonne Université et au sein du LORIA, avec le groupe Ethics@loria.

37. Voir : <http://www.ethique-et-tal.org/>.

Perspectives et projet de recherche

Sommaire

4.1	Un environnement structurant	79
4.2	Produire (de manière éthique) des ressources langagières pour le TAL	80
4.3	Mener la réflexion éthique en TAL et en IA	83
4.4	Croiser les chemins	87

4.1 Un environnement structurant

Je suis devenue, en 2019, chercheuse associée au LORIA. J’ai obtenu en 2021 que soit signée une convention entre la Sorbonne et le laboratoire, qui me permet de faire ma recherche officiellement au LORIA, dans l’équipe Sémagramme. En effet, j’oriente désormais mes activités de recherche sur le sujet de l’« éthique et IA » et le creuset lorrain, avec en particulier le projet OLKI (Open Language and Knowledge for citizens) de l’Université de Lorraine¹ dans sa nouvelle version et la variété des recherches menées au LORIA, est particulièrement adapté pour cela.

Je participe actuellement à trois projets de recherche financés (voir Figure 4.1), un projet européen ICT 38, AI-Proficient, porté par l’Université de Lorraine (Benôit Iung, CRAN), et deux projets ANR, CODEINE, porté par Aurélie Névéol (LISN) et Autogramm, porté par Sylvain Kahane (Paris Nanterre). Je suis responsable scientifique pour le Loria de CODEINE et *Project Ethics Officer* (responsable de l’éthique) du projet AI-Proficient. Ma participation à Autogramm sera moins centrale, j’y apporterai mon expertise en annotation manuelle de corpus et en évaluation. Ces projets, en particulier les responsabilités et encadrements associés, vont très largement structurer ma recherche dans les quatre ans à venir.

Par ailleurs, ma charge de co-présidente du comité d’éthique d’ACL impacte d’ores et déjà ma vie de chercheuse et va sans doute l’influencer de plus en plus au cours des cinq années de mon mandat.

1. Voir : <https://lue.univ-lorraine.fr/fr/open-language-and-knowledge-citizens-olki>.

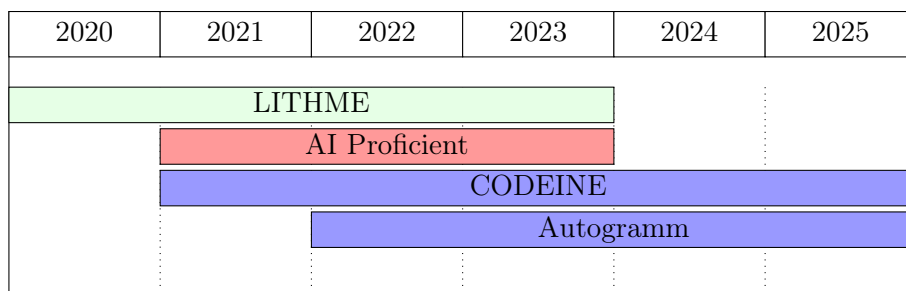


FIGURE 4.1 – Projets en cours (en vert les actions COST, en rouge les projets européens, en bleu les projets ANR).

4.2 Produire (de manière éthique) des ressources langagières pour le TAL

Il est pour moi fondamental d’ancrer ma réflexion éthique dans une pratique du domaine, je compte donc continuer à travailler dans la production de ressources langagières de qualité pour le TAL. Je développe actuellement cette activité selon trois axes principaux, en m’appuyant sur les sciences participatives et les jeux ayant un but.

4.2.1 Des ressources sémantiques libres pour le français

Méthodes formelles

L’une des thématiques de recherche principales de l’équipe Sémagramme est la sémantique. Il existe aujourd’hui de nombreux formalismes d’annotation en sémantique (AMR, DMRS, UCCA, DRS, etc.), mais très peu de ressources pour le français. Nous avons commencé à travailler sur les différents formalismes existants, afin de produire un corpus de taille réduite du français annoté dans plusieurs d’entre eux, à l’image du corpus utilisé dans la *shared task* MRP 2019 (*Cross-Framework Meaning Representation Parsing*). En parallèle, nous avons participé à la *shared task* ISA 17 (Interoperable Semantic Annotation), ce qui nous a permis de commencer à travailler sur l’annotation des quantifieurs en anglais et a donné lieu à une publication (Amblard et al., 2021).

Je compte participer aux collaborations mises en place par mon équipe avec des collègues à l’international, notamment Johan Bos (Univ. of Groningen), pour créer une dynamique visant à enrichir les bases multilingues existantes comme le *parallel*

4.2 Produire (de manière éthique) des ressources langagières pour le TAL

meaning bank (PMB). Mon apport pourrait consister à utiliser pour cela une plateforme ludifiée, librement adaptable pour n'importe quelle langue. Johan Bos a déjà expérimenté ce type de plateforme avec WordRobe et je pense qu'il est possible de trouver des solutions souples, adaptables pour plusieurs langues, en fonction des ressources langagières disponibles, qui sont souvent hétérogènes. Cette dimension du projet pourrait d'ailleurs intéresser également Mathieu Constant (Université de Lorraine, ATILF), qui travaille sur les liens entre ressources langagières, et Alain Polguère (Université de Lorraine, ATILF), qui a créé un réseau lexical sémantique du français de grande précision (RL-fr).

Plongements de graphes

Je co-encadre depuis septembre 2021 avec Mathieu Constant la thèse de Heesoo Choi, financée par l'école doctorale Sociétés, Langages, Temps, Connaissances (Université de Lorraine). Le sujet de la thèse porte sur la mise en correspondance de ressources linguistiques, en particulier sémantiques, pour le français. Nous souhaitons dans ce cadre expérimenter l'alignement automatique de ressources lexicales, afin de produire une ressource sémantique dynamique, librement disponible pour le français.

Relier entre elles deux ressources consiste à lier leurs entrées respectives. Cette tâche s'est révélée extrêmement complexe et impossible à réaliser avec des heuristiques simples sans perte d'informations (Guillaume et al., 2014). Nous nous proposons d'utiliser des plongements de graphes, afin de calculer, pour chaque entrée, un profil linguistique, puis de mesurer la proximité entre chaque paire d'entrées en comparant leurs profils. L'évaluation de notre méthode sera réalisée de manière intrinsèque (à l'aide de références établies manuellement) et extrinsèque (sur des tâches de TAL comme la résolution d'ambiguïtés lexicales). H. Choi a commencé à travailler sur le réseau lexical de *JeuxDeMots* (Rezo), et le Réseau lexical du français (RL-fr), deux ressources complémentaires très riches sémantiquement. Nous allons dans un premier temps nous concentrer sur quelques entrées bien choisies, dont certaines sélectionnées pour leur richesse, comme l'entrée « soleil » dans le RL-fr, et d'autres plus standards. Nous espérons ainsi bâtir une première référence qui nous permettra d'évaluer nos résultats.

Ce projet est très ambitieux, mais des résultats même partiels permettraient d'offrir un accès inédit en français à une ressource sémantique de grande qualité et à couverture large.

4.2.2 Des ressources synthétiques pour préserver le secret médical

Le projet CODEINE, porté par Aurélie Névéol (LISN) en collaboration avec le CEA LIST et l'INSERM, vise à produire des corpus médicaux synthétiques, afin de pouvoir les partager et produire des outils de TAL à partir de ceux-ci.

Je suis responsable scientifique du projet pour le LORIA et, en collaboration avec Bruno Guillaume, de la partie sciences participatives. En effet, nous prévoyons d'utiliser des jeux ayant un but pour i) valider les corpus synthétiques et ii) produire des annotations temporelles sur ceux-ci. Ce projet donnera lieu au recrutement à partir de l'automne 2022, d'un ingénieur au LORIA pour le développement de la plateforme de jeux.

Dans le cadre de ce même projet, je co-encadre depuis septembre 2021 la thèse de Nicolas Hiebel avec Aurélie Névéol et Olivier Ferret (CEA-LIST). Cette thèse est la continuité du mémoire de Master 2 dont j'étais l'encadrante et elle vise la création éthique de données textuelles artificielles, avec une application au domaine clinique. Une large partie du travail porte actuellement sur l'utilisation et l'adaptation de modèles de *deep learning* pour la tâche, en particulier des réseaux siamois, dont **SentenceBert** (Reimers and Gurevych, 2019). Cela me permet de monter en compétences sur ces sujets avec des collègues plus spécialisés, tout en apportant mon expertise des données et de l'éthique. Les premiers travaux menés dans ce cadre ont donné lieu à la création du premier corpus (1 000 phrases) de similarité textuelle dans le domaine clinique pour le français. Nous avons utilisé ce corpus pour adapter **SentenceBert**, apportant ainsi une amélioration significative aux résultats de l'outil sur des textes cliniques. Ce travail a fait l'objet d'une publication à LREC (Hiebel et al., 2022).

La suite du travail va consister à rechercher des moyens d'améliorer les performances des outils, en l'absence de grandes masses de données disponibles, tout en s'assurant de l'étanchéité du processus : les données médicales authentiques ne doivent en aucun cas être récupérables. Or, assurer la protection de la vie privée tout en permettant la création de données véritablement utilisables reste à l'heure actuelle un vrai défi (Stadler et al., 2021).

4.2.3 Des ressources pour évaluer les biais dans les modèles

Nous avons utilisé la plateforme de sciences participatives LanguageArc pour collecter des phrases porteuses de biais stéréotypés en français et valider nos traductions et nos choix de catégories de biais (Névéol et al., 2022). Nous avons mené une analyse détaillée des résultats en termes de participation et de qualité produite (Fort et al., 2022) cette première expérience montre d'ores et déjà l'intérêt des sciences participatives pour la réalisation de ce type de tâche. Nous sommes en

train d’élargir le corpus French CrowS-Pairs à d’autres langues, en collaboration avec des collègues et des étudiants originaires de différents pays : Margot Mieskes (allemand), Claudia Borg (maltais), Sergio Zanutto (italien) et Wolfgang Sebastian Schmeisser Nieto (espagnol).

Un autre axe que je compte approfondir est celui de l’évaluation des biais dans les modèles de type génératif. Des expériences bien conçues ont été menées pour évaluer les performances de ces modèles dans des conditions proches du réel, en comparant les résultats obtenus sur certaines tâches par des humains seuls et ceux produits par des humains assistés par un modèle (Casares et al., 2022). Je compte m’en inspirer pour concevoir une expérience d’évaluation des biais stéréotypés produits par ces modèles.

4.3 Mener la réflexion éthique en TAL et en IA

4.3.1 Favoriser la réflexion sur la recherche en TAL

Comité d’éthique d’ACL

Le travail que nous réalisons dans le cadre du comité d’éthique d’ACL comprend une part de questionnements, donc de recherche, et une part de pédagogie. Notre priorité est de mener le travail d’analyse de l’enquête que nous avons lancée début 2022 dans la communauté, afin de mieux comprendre où en est le domaine du point de vue de l’éthique et quelle est la perception du rôle de ce comité.

Parmi les nombreuses réflexions que nous allons être amenés à porter, il me semble que le rôle d’ArXiv doit être abordé en priorité. En effet, les citations en provenance d’ArXiv d’articles non publiés sont devenues monnaie courante aujourd’hui, ce qui a de nombreuses conséquences, dont la première est la remise en cause de l’évaluation par les pairs. Si celle-ci présente de nombreux écueils, sur lesquels il faut travailler (raison pour laquelle j’ai proposé avec des collègues des tutoriels sur la relecture à ACL 2020, EACL 2021 et TALN 2021), elle reste cependant la méthode d’évaluation de la recherche la moins biaisée lorsqu’elle est réalisée en double aveugle. ArXiv est également un extraordinaire accélérateur de la recherche, puisqu’un article peut être cité dès l’écriture terminée, sans passer par la phase parfois longue de publication. Non seulement cela entraîne des demandes ridicules de relecteurs qui exigent qu’on prenne en compte dans l’état de l’art un article publié quelques jours avant sur ArXiv, surtout, cela limite encore le temps de réflexion sur la recherche qu’on vient de mener.

Un autre sujet prioritaire est celui de la prise en compte de l’impact environnemental de nos recherches lors de la relecture, en particulier par les relecteurs éthiques. Ce sujet est complexe, car si de nombreux outils existent, aucun ne fait

encore consensus (Bannour et al., 2021).

Une réflexion à plus long terme concerne le choix de tracer ou non une ou des « lignes rouges » dans le domaine, autrement dit celui d’interdire ou non de publier sur certains sujets, comme la justice prédictive, voire de rejeter les recherches en provenance de pays à régimes autoritaires. Certains membres de la communauté poussent vers ce type de décision² alors que d’autres dénoncent une censure de la recherche, qui serait par définition neutre. S’il est évident pour moi que la recherche n’est absolument pas neutre, puisque son financement ne l’est pas, et qu’il est nécessaire de tracer certaines lignes rouges (la police et la justice prédictives en font partie), je suis contre l’interdiction de publication de collègues sur la base de leur pays de rattachement. Ces réflexions ont également lieu au niveau de l’IA en général et si les recommandations du groupe d’experts européen de haut niveau (High-Level Expert Group) sur l’IA ne comprennent aucune ligne rouge³, ce n’est pas tout à fait le cas dans le rapport de l’UNESCO qui a suivi⁴, qui définit certaines limites, notamment concernant les décisions de vie ou de mort. Le sujet est donc d’actualité et en pleine évolution et nous devons, en tant que spécialistes du TAL, participer à la discussion.

Surtout, nous devons créer des espaces d’échanges pour la communauté, afin que les discussions autour de ces sujets aient lieu avec le plus grand nombre.

Le comité d’éthique d’ACL rédige des rapports d’activité⁵. Nous nous réunissons entre présidents toutes les deux semaines et une fois par mois avec tous les membres. Les différences de fuseaux horaires et de rythmes de travail ne nous facilitent pas la tâche, mais nous tenons le cap de travailler en collectif, quitte à être moins efficaces. Nous avons fait ce choix, celui du temps nécessaire, en toute connaissance de cause, malgré les pressions pour aller plus vite.

Éthique déontologique

Je collabore depuis plusieurs mois avec Emily Bender, Margaret Mitchell (Hugging Face) et Emiel van Miltenburg (Université de Tilburg) sur l’analyse des sections « Ethical considerations » des articles d’ACL 2021. Les auteurs avaient en effet la possibilité d’ajouter une section de réflexion éthique à leur article, sans que celle-ci compte dans le nombre de pages maximum accepté et nous avons souhaité étudier les sujets abordés dans ces nouvelles sections. Nous avons réa-

2. C’est le cas notamment de Robert Munro : <https://towardsdatascience.com/research-on-machine-learning-for-disaster-response-b65f3e97c018>.

3. Voir : <https://www.euractiv.com/wp-content/uploads/sites/2/2018/12/AIHLEGDraftAIEthicsGuidelinespdf.pdf>.

4. Voir : https://unesdoc.unesco.org/ark:/48223/pf0000373434_fre.

5. Disponible ici en version préliminaire : https://www.aclweb.org/adminwiki/index.php?title=2022Q1_Reports:_Ethics_Committee_Co-chairs-Draft.

lisé une première phase d’annotation manuelle libre, qui nous a permis d’élaborer des catégories représentatives et les définitions associées (ce qui correspond à un guide d’annotation grossier). Nous réalisons en ce moment une annotation fine, avec identification des segments de texte correspondant à la catégorie identifiée. Il nous reste encore beaucoup de travail pour stabiliser les catégories et évaluer la cohérence de nos annotations (accord inter-annotateurs), avant de pouvoir procéder à une analyse sérieuse. Il ressort cependant de nos premiers résultats que ces sections de réflexion éthique sont très hétérogènes, du fait sans doute de leur nouveauté. Malgré certains excès (utilisation de la section pour étendre l’article sans pour autant porter de réflexion éthique), les auteurs se sont appropriés cet espace et ont dans l’ensemble joué le jeu. Ces résultats très préliminaires restent bien entendu à consolider et à creuser, mais ils semblent indiquer que le domaine du TAL s’approprie le sujet.

Enfin, je co-encadre, avec Maxime Amblard, Fanny Ducel dans le cadre de son mémoire de M1 (de janvier à septembre 2022) sur un travail qui vise à étudier la force des revendications (*claims*) dans les articles d’ACL. Étant donnée la difficulté de l’annotation manuelle des revendications, nous avons décidé d’utiliser une méthode non-supervisée à base de clustering pour analyser les résumés (*abstracts*), les introductions, les corps et conclusions d’articles de la conférence ACL 2021. Il en ressort pour l’instant que les auteurs ont tendance à émettre des revendications de plus en plus certaines au fil de l’article, ou à rester à un degré stable. Nous avons ensuite recoupé ces informations avec les données concernant les auteurs : genre, continent et institutions. Les résultats montrent qu’une majorité d’articles est encore écrite par des hommes, affiliés à l’Amérique du Nord, l’Asie ou l’Europe, et que, dans plus d’un tiers des cas, ils proviennent d’une institution renommée. Ces catégories les plus représentées sont également celles qui émettent le plus de revendications. Nous allons approfondir ce travail dans le but de le publier, mais ces résultats sont d’ores et déjà intéressants.

Par ailleurs, j’ai contacté les auteurs de l’article cité en Section 3.3.3 sur l’influence des « Big Tech » ([Abdalla and Abdalla, 2021](#)), dans le but d’approfondir la recherche sur le sujet en la centrant sur le TAL. Le groupe de travail inclut actuellement Mohamed Abdalla (Université de Toronto, Canada), Saif Mohammad (National Research Council, Canada), Terry Lima Ruas et Jan Philip Wahle (Université de Wuppertal, Allemagne), Aurélie Névél et Fanny Ducel. Nous progressons rapidement et espérons publier nos résultats en 2023.

4.3.2 Participer à la dynamique « Éthique et IA » du LORIA

Éthique dès la conception

Je suis *Project Ethics Officer* du projet AI Proficient et j'encadre au LORIA sur le sujet un post-doctorant spécialiste de l'éthique des valeurs, Marc Anderson. Ce projet vise à utiliser l'intelligence artificielle pour améliorer les processus industriels, il implique donc de nombreuses entreprises, en particulier Continental et Ineos. Mon rôle consiste à mettre en œuvre une éthique dès la conception, ou *ethics by design*, afin de s'assurer que les risques éthiques sont pris en compte dès la définition des *use cases*, jusqu'à leur implantation en entreprise. Cela implique un suivi constant des différents groupes de travail, qui est assuré par M. Anderson, sous ma supervision.

Le projet a commencé fin 2020 et nous avons produit à six mois un premier livrable public⁶ très important pour le projet (*Legal and ethical requirements for human-machine interaction*), puisqu'il décrit notre fonctionnement et les recommandations que nous avons faites pour chaque *Use Case*. M. Anderson et moi-même avons publié deux articles de revue dans ce cadre, concernant d'une part nos réflexions et travaux sur l'éthique dans la conception (Anderson and Fort, 2022a) et d'autre part la représentation de l'implication de l'humain dans les applications d'IA (Anderson and Fort, 2022b). Cette collaboration interdisciplinaire est très riche pour nous deux. Il nous a fallu plusieurs mois pour apprendre à travailler ensemble et comprendre le projet, mais les discussions que nous avons sont très enrichissantes et stimulantes intellectuellement. J'ai souhaité prolonger cette collaboration et l'élargir à d'autres collègues, aussi bien au LORIA qu'à l'extérieur.

Ainsi, outre le CRAN et les partenaires du projet, le groupe éthique collabore avec l'INRS, en particulier avec V. Govaere, pour ce qui concerne la sécurité au travail. Les perspectives de recherche sont multiples, car le domaine en est encore à ses balbutiements et les questions éthiques sont nombreuses et peu explorées. En outre, la multidisciplinarité du groupe (Marc Anderson est un philosophe et Christophe Cerisara spécialiste en *machine learning*) lui donne une vision unique sur ces questions.

Animer la réflexion éthique au LORIA et au niveau national

Par ailleurs, je suis à l'origine du groupe de travail Ethics@loria, qui réunit des chercheurs intéressés par le sujet (Maxime Amblard, Armelle Brun, Slim Ouni, Abdessamad Imine, Mathieu D'Aquin, Marc Anderson) et Aurore Coince, qui nous

6. Le document est accessible ici : https://ai-proficient.eu/wp-content/uploads/2021/09/D1.2-Legal-and-ethical-requirements-for-human-machine-interaction_v1.0.pdf.

aide à organiser et animer le collectif. Nous avons commencé à travailler ensemble en novembre 2021 et nous nous réunissons une fois par mois. J’ai proposé d’organiser un événement pour les doctorants du laboratoire, à la manière de l’atelier *Re-coding Black Mirror* organisé entre autres par Mathieu D’Aquin : les doctorants seront invités à travailler en groupes pour produire une dystopie autour d’un sujet proche de leurs sujets de thèses. Comme dans les hackathons, un prix pourrait être décerné par les participants à la production la plus originale et les projets les plus aboutis pourraient être publiés, par exemple sous forme de BD. Cet événement constituera un espace de réflexion éthique ludique, qui, s’adressant à de jeunes chercheurs, permettra de faire progresser le sujet de manière efficace dans nos domaines. Grâce à Aurore Coince, nous avons obtenu de l’école doctorale qu’il fasse partie de la formation des doctorants. Il aura lieu sur deux jours, début novembre 2022⁷.

Enfin, j’ai commencé à lancer une dynamique autour de l’éthique dans le cadre du GDR LIFT (Linguistique informatique, formelle et de terrain). Je compte mettre en place des rencontres croisées entre talistes et linguistes, afin de provoquer des échanges sur les pratiques et les expériences liées à l’éthique. Mon idée est d’organiser un atelier à TALN, avec des interventions invitées de linguistes de terrain, puis un atelier dans un cadre linguistique (à définir) avec des intervenants du TAL. Il me semble que les deux communautés bénéficieraient grandement de tels échanges et j’espère ainsi initier des collaborations. J’ai d’ailleurs été invitée à présenter une conférence sur l’éthique dans le cadre des journées conjointes des GDR TAL et LIFT, en novembre 2022, ce qui montre qu’il y a un besoin.

4.4 Croiser les chemins

Les possibilités de dystopies générées par le TAL sont trop nombreuses et se réalisent trop souvent pour que nous détournions le regard. Pour devenir des moteurs de progrès sociaux, le TAL et l’IA en général doivent s’ouvrir, non seulement vers d’autres disciplines, comme la philosophie et la sociologie, mais également vers les citoyens, afin de donner à ceux-ci les moyens d’articuler une réflexion sur le sujet et de participer aux décisions.

Mes chemins de recherche m’ont justement menée à ce croisement. Je compte dans les années à venir construire des projets visant à impliquer les citoyens et des collègues d’autres disciplines dans l’analyse, afin de proposer une manière différente d’envisager les recherches en TAL.

7. Voir : <https://lstu.fr/ethics-through-dystopia>.

Bibliographie

- (1979). The belmont report : Ethical principles and guidelines for the protection of human subjects of research.
- (2014). *Fairness, accountability, and transparency in machine learning (FATML)*. S. Barocas and M. Hardt.
- Abdalla, M. and Abdalla, M. (2021). The grey hoodie project : Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 287–297, New York, NY, USA. Association for Computing Machinery.
- Abeillé, A., Clément, L., and Toussanel, F. (2003). Building a treebank for French. In Abeillé, A., editor, *Treebanks*, pages 165 –187. Kluwer, Dordrecht.
- Adda, G. and Mariani, J. (2010). Language resources and amazon mechanical turk : legal, ethical and other issues. In *Proceedings of the Legal Issues for Sharing Language Resources workshop in International Conference on Language Resources and Evaluation (LREC)*, La Valette, Malte. European Language Resources Association (ELRA).
- Alves, D. V. A. (2018). Une foule de données créée par la foule est-elle suffisante ? évaluation extrinsèque des annotations en syntaxe de dépendances produites gr mbilingo. Master’s thesis, Master 1 Spécialité Langue et Informatique – Sorbonne Université. Encadré par Karën Fort et Bruno Guillaume.
- Amblard, M. (2016). Pour un TAL responsable. *Revue TAL*, 57(2) :21 – 45.
- Amblard, M., Fort, K., Guillaume, B., de Groote, P., Li, C., Ludmann, P., Musiol, M., Pavlova, S., Perrier, G., and Pogodalla, S. (2021). The annotators did not agree on some of the guidelines examples. In *Proceedings of the IWCS workshop ISA-17, the Seventeenth Workshop on Interoperable Semantic Annotation*, Groningen, Netherlands.
- Amblard, M., Fort, K., Musiol, M., and Rebuschi, M. (2014). L’impossibilité de l’anonymat dans le cadre de l’analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

BIBLIOGRAPHIE

- Anderson, M. and Fort, K. (2022a). From the ground up : developing a practical ethical methodology for integrating ai into industry. *AI & Society*.
- Anderson, M. and Fort, K. (2022b). Human where ? a new scale defining human involvement in technology communities from an ethical standpoint. *IRIE - International Review of Information Ethics*, 31.
- 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*.
- Aristote (-350). *Éthique à Nicomaque*. Wikisource.
- Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4) :555–596.
- Badouard, R., Barthe, Y., Cazalis, F., Chavalarias, D., Chlous, F., de Rosnay, M. D., Fort, K., Fourniau, J.-M., Guillaud, D., Julliard, R., Marec, J. L., Mabi, C., Meyer, M., Pitrou, P., Prevot, A.-C., Roturier, C., and Tsoukias, A. (2016). Position paper. Technical report, GPRO Sciences participatives (dir. Sandra Laugier). Alliance ATHENA.
- Bannour, N., Ghannay, S., Névéol, A., and Ligozat, A.-L. (2021). Evaluating the carbon footprint of NLP methods : a survey and analysis of existing tools. In *EMNLP, Workshop SustaiNLP*, Punta Cana, Dominican Republic.
- Barbaresi, A. and Lejeune, G. (2020). Que recèlent les données textuelles issues du web ? In *Actes de l’atelier Éthique et TRaitemeNt Automatique des Langues@JEP-TALN-RECITAL 2020. 2e atelier ÉThique et tRaitemeNt Automatique des Langues (ETeRNAL)*, pages 19–28, Nancy, France. Association pour le Traitement Automatique des Langues. What do text data from the Web have to hide ?
- Bartle, R. (1996). Hearts, clubs, diamonds, spades : Players who suit MUDs. *The Journal of Virtual Environments*, 1(1).
- Baude, O., Blanche-Benveniste, C., Calas, M.-F., Cappeau, P., Cordereix, P., Goury, L., Jacobson, M., De Lamberterie, I., Marchello-Nizia, C., and Mondada, L. (2006). *Corpus oraux, guide des bonnes pratiques 2006*. CNRS Editions, Presses Universitaires Orléans.
- Begue, H. (2019). Développement de ressources langagières et d’outils de TAL pour le créole mauricien. Mémoire de Master 1 rbonne Université.

- Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3).
- Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing : Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6 :587–604.
- 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.
- Bentham, J. (1780). *An Introduction to the Principles of Morals and Legislation*. Wikisource.
- Bernard, L., Besombes, C., Boula De Mareüil, P., Chupin, L., Dagorne, E., Delannoy, M., Desainte-Catherine, M., Dosseur, B., Drouin, V., Echassoux, A., Fort, K., Guillaud, D., Girard, J.-P., Ilien, G., Julliard, R., Laborde, D., Lemaire, F., Lheureux, R., L’Her, G., Mathieu, Y., Pellerin, G., Puig, V., Quach, C., Severo, M., Sinclair, P. f., Siret, D., and Vurpillot, D. (2019). Recherche culturelle et sciences participatives PARTICIP-ARC. Research report, Muséum national d’Histoire naturelle.
- Bernhard, D. and Ligozat, A.-L. (2013). Es esch fäscht wie Ditsch, oder net ? étiquetage morphosyntaxique de l’alsacien en passant par l’allemand. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d’Europe*, pages 209–220, Les Sables d’Olonne, France.
- 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., Daumé III, H., and Wallach, H. (2020). Language (technology) is power : A critical survey of "bias" in nlp. In *ACL*.
- Bonastre, J., Bimbot, F., Jean Boë, L., Campbell, J. P., Reynolds, D. A., and Magrin-chagnolleau, I. (2003). Person authentication by voice : A need for caution. In *Proc. of Eurospeech 03*.
- Bonastre, J.-F. (2020). 1990-2020 : retours sur 30 ans d’échanges autour de l’identification de voix en milieu judiciaire. In *Actes de l’atelier Éthique et Traitement Automatique des Langues@JEP-TALN-RECITAL 2020. 2e atelier ÉThique et*

BIBLIOGRAPHIE

- tRaitement Automatique des Langues (ETeRNAL)*, pages 38–47, Nancy, France. Association pour le Traitement Automatique des Langues. 1990-2020 : A look back at 30 years of discussions on voice identification in the judicial system.
- Boudin, F. (2013). Taln archives : une archive numérique francophone des articles de recherche en traitement automatique de la langue. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles*, pages 507–514, Les Sables d’Olonne, France. Association pour le Traitement Automatique des Langues.
- Bouquiaux, L. and Thomas, J. (1971). *Enquête et description des langues à tradition orale. : l’enquête de terrain et l’analyse grammaticale*, volume I. Société d’études linguistiques et anthropologiques de France, Paris, 2nd edition 1976 edition.
- Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Burnard, L. (2005). Metadata for corpus work. *Developing linguistic corpora : A guide to good practice*, pages 30–46.
- Caillois, R. (1958). *Les Jeux et les Hommes : le masque et le vertige*. Folio/Essais. Gallimard, Paris, ition revue et augment1967] edition.
- Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., and de la Clergerie, E. (2014). Deep syntax annotation of the sequoia french treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.
- Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l’adaptation d’analyseur par pont lexical. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.
- Casares, P. A. M., Loe, B. S., Burden, J., O’hEigearthaigh, S., and Hernez-Orallo, J. (2022). How general-purpose is a language model ? usefulness and safety with human prompters in the wild. In *Proceedings of AAAI 2022*.
- Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources : Successes and limitations of the approach. In Gurevych, I. and Kim, J., editors, *The People’s Web Meets NLP, Theory and Applications of Natural Language Processing*, pages 3–44. Springer Berlin Heidelberg.

- Chamberlain, J., Kruschwitz, U., and Poesio, M. (2009a). Constructing an anaphorically annotated corpus with non-experts : assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, People's Web '09, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase Detectives : a web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz, Autriche.
- Chamberlain, J., Poesio, M., and Kruschwitz, U. (2009b). A new life for a dead parrot : Incentive structures in the phrase detectives game. In *Proceedings of WWW 2009*, Madrid, Espagne.
- Chupin, L. and Fort, K. (2019). Ouvrir le dédale des données des recherches myriadisées. *Culture et recherche*, 140.
- Cohen, K. B., Fort, K., Adda, G., Zhou, S., and Farri, D. (2016). Ethical Issues in Corpus Linguistics And Annotation : Pay Per Hit Does Not Affect Effective Hourly Rate For Linguistic Resource Development On Amazon Mechanical Turk. In *ETHics In Corpus collection, Annotation and Application workshop*, Proceedings of the ETHics In Corpus collection, Annotation and Application workshop, Portoroz, Slovenia.
- Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A. C., Beenen, M., Salesin, D., Baker, D., and Popović, Z. (2010). The challenge of designing scientific discovery games. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, pages 40–47, New York, NY, USA. ACM.
- Couillault, A. and Fort, K. (2013). Charte éthique et big data : parce que mon corpus le vaut bien ! In *Actes de colloque international Corpus et Outils en Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasbourg, France.
- Couillault, A., Fort, K., Adda, G., and De Mazancourt, H. (2014). Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter. In *International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Islande.
- Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009). Complex linguistic annotation - no easy way out ! a case from bangla and hindi POS labeling tasks. In *Proceedings of the third ACL Linguistic Annotation Workshop*, Singapour.

BIBLIOGRAPHIE

- de Castilho, R. E., Biemann, C., Gurevych, I., and Yimam, S. M. (2014). Webanno : a flexible, web-based annotation tool for clarin. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*, page online, Utrecht, Netherlands. CLARIN ERIC. Extended abstract.
- de Chalendar, G. (2014). Traitement automatique des langues, biens communs informationnels et industries de la langue. In *Journée ATALA éthique et TAL*, Paris, France.
- de Mazancourt, H., Couillault, A., and Recourcé, G. (2014). L’anonymisation, pierre d’achoppement pour le traitement automatique des courriels. In *Journée ATALA éthique et TAL*, Paris, France.
- Denis, P. and Sagot, B. (2010). Exploitation d’une ressource lexicale pour la construction d’un étiqueteur morphosyntaxique état-de-l’art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada.
- Desrosières, A. (1989). Comment faire des choses qui tiennent : histoire sociale et statistique. *Histoire & Mesure*, 4(4) :225–242.
- Desrosières, A. (2001). Entre réalisme métrologique et conventions d’équivalence : les ambiguïtés de la sociologie quantitative. *Genèses*, 2(43) :112–127.
- Desrosières, A. (2008). *Pour une sociologie historique de la quantification : L’Argument statistique*. Presses de l’école des Mines de Paris.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dittrich, D. and Kenneally, E. (2012). The Menlo Report : Ethical Principles Guiding Information and Communication Technology Research. Technical report, U.S. Department of Homeland Security.
- Drugan, J. and Babych, B. (2010). Shared resources, shared values ? ethical implications of sharing translation resources. In *Proceedings of the Second Joint EM+/CNGL Workshop : Bringing MT to the User : Research on Integrating MT in the Translation Industry*, pages 3–10, Denver, Colorado, USA. Association for Machine Translation in the Americas.

- Enguehard, C. and Mangeot, M. (2014). Favorisons la diversité linguistique en tal. In *Journée ATALA éthique et TAL*, Paris, France.
- Eshkol-Taravella, I., Kanaan-Caillol, L., Baude, O., Dugua, C., and Maurel, D. (2014). Procédure d’anonymisation et traitement automatique : l’expérience d’ESLO. In *Journée ATALA éthique et TAL*, Paris, France.
- Fenouillet, F., Kaplan, J., and Yennek, N. (2009). Serious games et motivation. In *4ème Conférence francophone sur les Environnements Informatiques pour l’Apprentissage Humain (EIAH’09), vol. Actes de l’Atelier "Jeux Sérieux : conception et usages"*, pages 41–52, Le Mans, France.
- Fiumara, J., Cieri, C., Wright, J., and Liberman, M. (2020). LanguageARC : Developing language resources through citizen linguistics. In *Proceedings of the LREC 2020 Workshop on “Citizen Linguistics in Language Resource Development”*, pages 1–6, Marseille, France. European Language Resources Association.
- Floridi, L. (2013). *The Ethics of Information*. Oxford University Press, Oxford, England, UK.
- Floridi, L. and Chiriatti, M. (2020). Gpt-3 : Its nature, scope, limits, and consequences. *Minds & Machines*, 30 :681–694.
- Fort, K. (2012). *Les ressources annotées, un enjeu pour l’analyse de contenu : vers une méthodologie de l’annotation manuelle de corpus*. PhD thesis, Université Paris XIII, LIPN, INIST-CNRS.
- Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing*. Focus series. ISTE Wiley.
- Fort, K. (2017). Experts ou (foule de) non-experts ? la question de l’expertise des annotateurs vue de la myriadisation (crowdsourcing). *CORELA (cognition, représentation, langage)*, page 12 p.
- 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., Adda, G., Sagot, B., Mariani, J., and Couillault, A. (2014a). Crowdsourcing for Language Resource Development : Criticisms About Amazon Mechanical Turk Overpowering Use. In Vetulani, Z. and Mariani, J., editors, *Human Language Technology Challenges for Computer Science and Linguistics*, pages 303–314. Springer International Publishing.
- Fort, K. and Amblard, M. (2018). Éthique et traitement automatique des langues. In *Journée éthique et intelligence artificielle*, Nancy, France.

BIBLIOGRAPHIE

- Fort, K. and Couillault, A. (2016). Yes, We Care! Results of the Ethics and Natural Language Processing Surveys. In *international Language Resources and Evaluation Conference (LREC) 2016*, Proceedings of the international Language Resources and Evaluation Conference (LREC) 2016, Portoroz, Slovenia.
- Fort, K. and Guillaume, B. (2019). Les jeux ayant un but : des sciences participatives? *Culture et recherche*, 140.
- Fort, K., Guillaume, B., and Chastant, H. (2014b). Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. In *Gamification for Information Retrieval (GamifIR'14) Workshop*, Amsterdam, Pays-Bas.
- Fort, K., Guillaume, B., Constant, M., Lefèbvre, N., and Pilatte, Y.-A. (2018a). "Fingers in the Nose" : Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform. In *LAW-MWE-CxG 2018 - COLING 2018 Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 207 – 213, Santa Fe, United States.
- Fort, K., Guillaume, B., and Lefèbvre, N. (2017a). Who wants to play Zombie? A survey of the players on ZOMBILINGO. In *Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, page 2, Valencia, Spain.
- Fort, K., Guillaume, B., Lefèbvre, N., Ramírez, L., Regnault, M., Collins, M., Gavrilova, O., and Kristanti, T. (2017b). Vers l'annotation par le jeu de corpus (plus) complexes : le cas de la langue de spécialité. In *Traitement Automatique des Langues Naturelles (TALN)*, Orléans, France.
- Fort, K., Guillaume, B., Pilatte, Y.-A., Constant, M., and Lefre, N. (2020). Rigor mortis : Annotating mwes with a gamified platform. In *Proc. of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.
- Fort, K., Lafourcade, M., and Brun, N. L. (2018b). Cheap, fast and good! voting games with a purpose. In *Actes de l'atelier LREC Games4NLP 2018*, Miyazaki, Japon.
- Fort, K., Nazarenko, A., and Rosset, S. (2012). Modeling the complexity of manual annotation tasks : a grid of analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 895–910, Mumbai, Inde.

- Fort, K. and Névéol, A. (2018). Présence et représentation des femmes dans le traitement automatique des langues en france. In *Actes de l'atelier "Penser la Recherche en Informatique comme pouvant être Située, Multidisciplinaire Et Genrée" (PRISME-G)*.
- Fort, K., Névéol, A., Dupont, Y., and Bezançon, J. (2022). Use of a Citizen Science Platform for the Creation of a Language Resource to Study Bias in Language Models for French : a case study. In *2nd LREC Workshop on Novel Incentives in Data Collection from People*, Marseille, France.
- 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.
- 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.
- Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011). Managing the crowd : Towards a taxonomy of crowdsourcing processes. In *AMCIS 2011 Proceedings*.
- Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources : Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japon.
- Guillaume, B., Fort, K., Perrier, G., and Bédaride, P. (2014). Mapping the lexique des verbes du français (lexicon of french verbs) to a nlp lexicon using examples. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.
- Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Texte !*, vol. X(4).
- Hagendorff, T. (2020). The ethics of ai ethics : An evaluation of guidelines. *Minds & Machines*, 30 :99–120.
- Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In *CHI 2018*, Montreal, QC, Canada.
- Hara, K., Adams, A., Milland, K., Savage, S., Hanrahan, B. V., Bigham, J. P., and Callison-Burch, C. (2019). Worker demographics and earnings on amazon

BIBLIOGRAPHIE

- mechanical turk : An exploratory analysis. CHI EA '19, pages 1–6, New York, NY, USA. Association for Computing Machinery.
- Hiebel, N., Ferret, O., Fort, K., and Névél, A. (2022). CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives. In *LREC 2022 - International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.
- Houllier, F. and Merilhou-Goudard, J.-B. (2016). Les sciences participatives en France. Other. Ce rapport est présenté en 3 livrets : états des lieux - bonnes pratiques - recommandations Ont également contribué à la réflexion et à la rédaction : Mathieu Andro (annexe10), François Charbonnel (annexes 3, 4, 5, 6), Jean-Philippe Cointet (annexe 3), Pascale Frey-Klett (livrets 2, 3), Pierre-Benoit Joly (livrets 1, 3, annexes 3, 4, 5, 6, 8), Hugues Leiser (annexe 8), et Muriel Mambrini-Doudet (livrets 2, 3, annexes 5, 6, 7, 8) Ont également contribué à la réflexion et à la relecture : Odile Hologne, Jean-François Launay, Olivier Le Gall, Jean Masson, Nathalie Morcrette, Jean-Luc Pujol et Christophe Roturier.
- Hovy, D. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8) :e12432.
- 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.
- Ipeirotis, P. (2010). Analyzing the amazon mechanical turk marketplace. *ACM Crossroads*, 17 :16–21.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Kant, E. (1781). *Critique de la raison pure*. Wikisource.
- Kant, E. (1788). *Critique de la raison pratique*.
- Kenny, D. (2011). The ethics of machine translation. In *New Zealand Society of Translators and Interpreters Annual Conference 2011*, Auckland, New Zealand.
- Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., et al. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10) :1175–1177.

- Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform : Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title : The 27th International Conference on Computational Linguistics (COLING 2018).
- Lafourcade, M. (2007). Making people play for lexical acquisition. In *Proceedings of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thaïlande.
- Lafourcade, M., Brun, N. L., and Joubert, A. (2015a). *Games with a Purpose (GWAPS)*. Wiley-ISTWiley-ISTE.
- Lafourcade, M. and Fort, K. (2014). Propa-l : a semantic filtering service from a lexical network created using games with a purpose. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.
- Lafourcade, M., Joubert, A., and Brun, N. (2018). The jeuxdemots project is 10 years old : What we have learned. In *Proc. of the LREC workshop Games4NLP 2018*, Miyazaki, Japan.
- Lafourcade, M., Le Brun, N., and Joubert, A. (2015b). Collecting and evaluating lexical polarity with a game with a purpose. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria.
- Lafourcade, M., Le Brun, N., and Joubert, A. (2016). Construire un lexique de sentiments par crowdsourcing et propagation. In *Proc. of Traitement Automatique des Langues Naturelles*, Paris, France.
- Lafourcade, M., Le Brun, N., and Zampa, V. (2014). Crowdsourcing word-color associations. In *Proc. of the International Conference on Application of Natural Language to Information Systems (NLDB)*, Montpellier, France.
- Lafourcade, M. and Lebrun, N. (2014). éthique et construction collaborative de données lexicales par des gwaps. Journée d’étude ATALA Éthique et TAL.
- Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT : Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

BIBLIOGRAPHIE

- Leech, G. (1991). The state of the art in corpus linguistics. *English Corpus Linguistics : Linguistic Studies in Honour of Jan Svartvik*, pages 8–29.
- Leech, G. (1997). *Corpus annotation : Linguistic information from computer text corpora*, chapter Introducing corpus annotation, pages 1–18. Longman, Londres, Angleterre.
- Lefevre, 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.
- Lefevre-Halftermeyer, A., Govaere, V., Antoine, J.-Y., Allegre, W., Pouplin, S., Departe, J.-P., Slimani, S., and Spagnulo, A. (2016). Typologie des risques pour une analyse éthique de l'impact des technologies du TAL. *Revue TAL*, 57(2) :47–71.
- Leidner, J. L. and Plachouras, V. (2017). Ethical by design : Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.
- Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., Szalay, A., Andreescu, D., Murray, P., and Vandenberg, J. (2010). Galaxy Zoo 1 : data release of morphological classifications for nearly 900 000 galaxies*. *Monthly Notices of the Royal Astronomical Society*, 410(1) :166–178.
- Lion-Bouton, A., Grobol, L., Antoine, J.-Y., Billot, S., and Lefevre-Halftermeyer, A. (2020). Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ? In *Actes de l'atelier Ethique et TRaitemeNt Automatique des Langues@JEP-TALN-RECITAL 2020. 2e atelier ÉThique et tRaitemeNt Automatique des Langues (ETeRNAL)*, pages 10–18, Nancy, France. Association pour le Traitement Automatique des Langues. Do the standard scores of evaluation of coreference resolution constitute metrics ?
- Machado, A. (1912). *Campos de Castilla*, chapter Proverbios y cantares.
- Macklin, R. (1999). *Against Relativism : Cultural Diversity and the Search for Ethical Universals in Medicine*. Oxford University Press, New York.

- Magistry, P., Ligozat, A.-L., and Rosset, S. (2018). Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux. In *Conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France.
- Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2) :313–330.
- Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Massé, R. (2000). Les limites d’une approche essentialiste des ethnoéthiques : Pour un relativisme éthique critique. *Anthropologie et Sociétés*, Anthropologie, relativisme éthique et santé(24-2) :13–33.
- Mathet, Y. and Widlöcher, A. (2016). Évaluation des annotations : ses principes et ses pièges. *Revue TAL*, 57(2) :73–98.
- Mill, J. S. (1859). *Bentham*.
- Millour, A. (2019). Getting to Know the Speakers : a Survey of a Non-Standardized Language Digital Use. In *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland.
- Millour, A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. Theses, Sorbonne Université.
- Millour, A., Dupont, Y., Jouglar, A., and Fort, K. (2022). FENEC : un corpus à échantillons équilibrés pour l’évaluation des entités nommées en français. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Avignon, France.
- Millour, A. and Fort, K. (2018a). À l’écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. *Revue TAL*, numéro spécial traitement automatique des langues peu dotées, 59(3) :41–65.
- Millour, A. and Fort, K. (2018b). Krik : First steps into crowdsourcing pos tags for kréyòl gwadloupéyen. In Soria, C., Besacier, L., and Pretorius, L., editors, *Proceedings of the Eleventh International Conference on Language Resources and*

BIBLIOGRAPHIE

- Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Millour, A. and Fort, K. (2018c). Toward a lightweight solution for less-resourced languages : Creating a pos tagger for alsatian using voluntary crowdsourcing. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Millour, A. and Fort, K. (2019). Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling. In *RANLP*, pages 776 – 784, Varna, Bulgaria.
- Millour, A., Fort, K., Bernhard, D., and Steiblé, L. (2017). Vers une solution légère de production de données pour le tal : création d’un tagger de l’alsacien par crowdsourcing bénévole. In *Proc. of TALN - Traitement Automatique des Langues Naturelles*, pages 139–154, Orléans, France.
- Millour, A., Fort, K., and Magistry, P. (2020). Répliquer et étendre pour l’alsacien "étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux". In *Actes de l’atelier Ethique et TRaitemeNt Automatique des Langues@JEP-TALN-RECITAL 2020. 2e atelier ÉThique et tRai-temeNt Automatique des Langues (ETeRNAL)*, pages 29–37, Nancy, France. Association pour le Traitement Automatique des Langues. Replicating and extending for Alsatian : "POS tagging for low-resource languages by adapting word embeddings".
- Millour, A., Grace Araneta, M., Lazić Konjik, I., Raffone, A., Pilatte, Y.-A., and Fort, K. (2019). Katana and Grand Guru : a Game of the Lost Words (DEMO). In *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC’19)*, Poznań, Poland.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19*, pages 220–229, New York, NY, USA. Association for Computing Machinery.
- Moosavi, N. S., Fan, A., Shwartz, V., Glavaš, G., Joty, S., Wang, A., and Wolf, T., editors (2020). *Proceedings of SustaiNLP : Workshop on Simple and Efficient*

- Natural Language Processing*, Online. Association for Computational Linguistics.
- Moosavi, N. S., Gurevych, I., Fan, A., Wolf, T., Hou, Y., Marasović, A., and Ravi, S., editors (2021). *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, Virtual. Association for Computational Linguistics.
- Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs : A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Névél, A., Dupont, Y., Bezançon, J., and Fort, K. (2022). French crows-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Irlande.
- Névél, A., Fort, K., and Hwa, R. (2017). Report on EMNLP Reviewer Survey. Technical report, Association for computational linguistics.
- Neves, M. and Seva, J. (2019). An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 22(1) :146–163.
- Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser : A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13 :95–135.
- Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives : Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1) :3 :1–3 :44.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.
- Rosa, H. (2012). *Aliénation et accélération – vers une théorie critique de la modernité tardive*. Collection Théorie critique. La Découverte, Paris.

BIBLIOGRAPHIE

- Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., and Zweigenbaum, P. (2012). Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. In *Proceedings of the 6th Linguistic Annotation Workshop (LAW VI)*, pages 40–48, Jeju, République de Corée.
- Sagot, B., Fort, K., Adda, G., Mariani, J., and Lang, B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France. 12 pages.
- Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, pages 254–263, Waikiki, Honolulu, Hawaii.
- Stadler, T., Oprisanu, B., and Troncoso, C. (2021). Synthetic data – anonymisation groundhog day. ArXiv paper.
- Stengers, I. (2019). *Résister au désastre*. Collection Petite bibliothèque d'écologie populaire. Marseille.
- Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.
- Straka, M., Straková, J., and Hajic, J. (2019). Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing.
- 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.
- Stubbs, A. (2012). Developing specifications for light annotation tasks in the biomedical domain. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Istanbul, Turquie.
- Tisserant, G. and Lafourcade, M. (2015). Politit, du crowd-sourcing pour politiser le lexique. In *Proc. of Etudier le Web politique : Regards croisés Institut des Sciences de l'Homme*, Lyon, France.
- Tuite, K. (2014). Gwaps : Games with a problem. *Foundations of Digital Games 2014*.

- Urieli, A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail, France.
- von Ahn, L. (2006). Games with a purpose. *IEEE Computer Magazine*, pages 96–98.
- von Ahn, L. (2013). Duolingo : learn a language for free while helping to translate the web. In ACM, editor, *Proceedings of the 2013 international conference on Intelligent user interfaces (IUI '13)*, pages 1–2, New York, NY, USA.
- von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '04*, pages 319–326, New York, NY, USA. ACM.
- Wynne, M., editor (2005). *Developing Linguistic Corpora : a Guide to Good Practice*. Oxford : Oxbow Books.
- 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.
- Zimmer, M. (2010). "but the data is already public" : on the ethics of research in facebook. *Ethics and Information Technology*, 12 :313–325.