Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Portabilité linguistique des modèles de langage pré-appris appliqués à la tâche du dialogue humain-machine en français

Untel Trucmuche[1, 2] Unetelle Machinchose[1, 3]
(1) Lab, adresse, CP Ville, Pays
(2) Lab, adresse, CP Ville, Pays
(3) Lab, adresse, CP Ville, Pays
utrucmuche@lab.fr, umachinchose@adresse-academique.be

RÉSUMÉ

Dans cet article, nous proposons une étude de la portabilité linguistique des modèles de langage pré-appris (MLPs) appliqués à une tâche de dialogue à domaine ouvert. La langue cible (L_T) retenue dans cette étude est le français. Elle dispose de peu de ressources spécifiques pour la tâche considérée et nous permet de réaliser une évaluation humaine in situ. La langue source (L_S) est l'anglais qui concentre la majorité des travaux récents dans ce domaine. Construire des MLPs spécifiques pour chaque langue nécessite de collecter de nouveaux jeux de données et cela est coûteux. Ainsi, à partir des ressources disponibles en L_S et L_T, nous souhaitons évaluer les performances atteignables par un système de conversation en L_T. Trois approches sont proposées : TrainOnTarget où le corpus L_S est traduit vers L_T avant l'affinage du modèle, TestOnSource où un modèle L_S est couplé avec des modules de traduction au moment du décodage et TrainOnSourceAdaptOnTarget, qui utilise un MLP multilingue - ici BLOOM (BigScience Workshop, 2022) - avec l'architecture MAD-X Adapter (Pfeiffer *et al.*, 2020) pour apprendre la tâche en L_S et l'adapter à L_T. Les modèles sont évalués dans des conditions de dialogue oral et les stratégies sont comparées en termes de qualité perçue lors l'interaction.

ABSTRACT

Linguistic portability strategies for open-domain dialogue with pre-trained language models from high to low resource languages

In this paper we propose a study of linguistic portability of pre-trained language models (PLMs) for open-domain dialogue systems in a high-resource language. The target language (L_T) is simulated with French as it lacks task-specific resources and allows an in-situ human evaluation. The source language (L_S) is English which concentrates the majority of recent works. Building specific PLMs for each possible language supposes collecting new datasets and is costly. Hence, leveraging resources from both L_S and L_T, we assess the performance achievable in L_T with three approaches : TrainOnTarget where a L_S dataset is translated in L_T before finetuning, TestOnSource where a L_S model is coupled with translation modules at inference and the TrainOnSourceAdaptOnTarget, using a multilingual PLM - here BLOOM (BigScience Workshop, 2022) - with MAD-X Adapter architecture (Pfeiffer *et al.*, 2020) to learn the task in L_S and adapt it to L_T. Models are evaluated in spoken dialogue conditions with human and the strategies compared in terms of perceived interaction quality.

MOTS-CLÉS : Agent conversationnel, Transformers, Portabilité multilingue, Langue peu dotée.

KEYWORDS: Conversational agent, Transformers, Crosslingual portability, Low-resource language.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 148–156.

1 Introduction

Depuis l'apparition des modèles Transformers (Vaswani *et al.*, 2017), plusieurs variantes de MLPs ont été déployées dans le domaine du traitement automatique du langage. Les Transformers autorégressifs (utilisant le bloc décodeur) comme GPT (Radford & Narasimhan, 2018), BART (Lewis *et al.*, 2019) etc. se positionnent sur l'état de l'art pour de nombreuses tâches génératives, dont le dialogue à domaine ouvert. Mais pour cela les systèmes doivent développer certaines capacités humaines telles que l'empathie, et avoir un personnalité consistante durant l'interaction (Walker *et al.*, 2021). Dans cette optique, des corpus spécifiques ont été collectés en faisant interagir des humains. On peut citer par exemple : PersonaChat (Zhang *et al.*, 2018), Empathetic Dialogues (Rashkin *et al.*, 2019), Blended Skill Talk (Smith *et al.*, 2020) etc. sur lesquels les MLPs peuvent être affinés. La majorité des corpus disponibles sont en anglais, ou en chinois. L'absence de corpus d'apprentissage spécifiques du même type en français, empêche l'obtention directe de modèles similaires.

Dans ce travail, nous étudions les stratégies de portabilité des chatbots et des corpus d'une langue source (L_S, ici l'anglais) vers une langue cible (L_T, ici le français). En exploitant le maximum de ressources disponibles dans L_S et L_T (outils de traduction automatique neuronale (TAN), corpus et MLPs), nous avons mis en place et mené une évaluation humaine de différents systèmes obtenus par trois approches : TrainOnSource, TestOnTarget, TrainOnsourceAndAdaptOnTarget. Nous avons ensuite comparé ces modèles à un modèle de référence en L_S à savoir BlenderBot 1.0 (Roller *et al.*, 2020).

Pour cela, nous avons revisité les approches proposées pour la portabilité linguistique des modules de compréhension de la parole (Jabaian *et al.*, 2013; Lefèvre *et al.*, 2010) afin de les appliquer au cas des MLPs pour la conversation orale humain-machine. A notre connaissance, l'un des seuls travaux à avoir abordé la question du développement multilingue des ressources pour les chitchat bots basés sur les PLMs est (Lin *et al.*, 2020) qui propose des traductions en plusieurs langues du corpus PersonaChat et l'utilise pour apprendre des modèles en différentes langues. Pour la modélisation du dialogue, nous avons utilisé le même schéma d'apprentissage que celui proposé par (Wolf *et al.*, 2019). En plus de reprendre ce qu'ils ont fait avec un modèle de type GPT (Radford & Narasimhan, 2018), nous avons exploré l'application de la même approche en français et avec le modèle BLOOM (BigScience Workshop, 2022), un modèle multilingue en libre accès [1].

2 Stratégies pour la portabilité des systèmes de dialogues de l'anglais vers le français

Dans cette étude préliminaire, plutôt que de nous concentrer sur l'amélioration de la performance intrinsèque du dialogue, nous évaluons comment les données et les modèles de L_S peuvent être exploités pour développer des modèles conversationnels simples basés sur des MLPs en L_T.

Les approches *TestOnSource* et *TrainOnTarget* s'appuient sur l'utilisation de modules de TAN à différentes étapes. Pour cela, nous avons utilisé l'API `Google Translate` comme dans (Lin *et al.*, 2020), excellentes performances et facilité d'utilisation expliquent ce choix parmi d'autres.

1. Le modèle est disponible sur https ://huggingface.co/bigscience/bloom

FIGURE 1 – **TestOnSource**

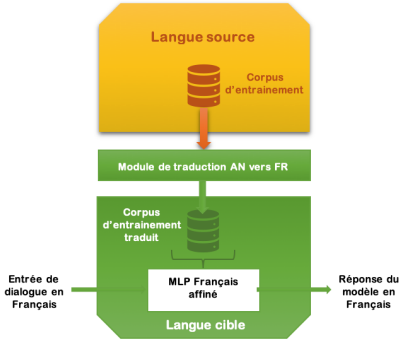

FIGURE 2 – **TrainOnTarget**

TestOnSource : L'approche consiste à utiliser les jeux de données, les modèles de dialogue et les MLPs disponibles en L_S et les combiner avec deux systèmes de TAN qui opèrent pendant les conversations bot-humain en L_T. Le premier (en orange dans la figure 1) traduit l'énoncé de l'utilisateur et le deuxième (en vert dans la figure 1) traduit la sortie du système. La disponibilité d'un grand nombre de ressources, notamment les modèles de dialogue à domaine ouvert en L_S, est un atout majeur pour cette approche. Par conséquent, il est intéressant d'évaluer les performances de ces systèmes sur des entrées traduites de L_T à L_S lors de l'inférence.

TrainOnTarget : Illustrée dans la figure 2, cette approche consiste à affiner les MLPs dans L_T (en vert) pour une tâche de chitchat sur un corpus traduit automatiquement depuis L_S (en jaune). Le français (L_T) dispose de quelques MLPs qui peuvent être utilisés comme base pour des modèles de dialogue. Cette approche suppose que les connaissances spécifiques à la langue, apprises par les MLPs en L_T, peuvent aider à gérer les échantillons bruités issus de la TAN.

TrainOnSourceAdaptOnTarget : Les approches précédentes s'appuient sur le fait qu'en dehors du chitchat, L_T est une langue dotée disposant de modèles de TAN et de MLPs ce qui n'est pas le cas de beaucoup de langues d'où l'idée d'utiliser des MLPs multilingues. Nous reproduisons l'architecture MAD-X (Pfeiffer *et al.*, 2020) pour le dialogue en utilisant **BLOOM** qui a des capacités de traduction et qui intègre une grande variété de langues peu dotées. Dans la figure 3, les flèches vides montrent la 1ère étape d'affinage des adaptateurs de tâche (sur les données de L_S) avec les adaptateurs de langue L_S gelés. Dans la 2ième étape (flèches pleines), les adaptateurs de langue (toujours gelés) passent de L_S à L_T et les mêmes adaptateurs de tâche sont affinés en utilisant peu de données L_T (ou des données traduites). En amont de ces étapes, les adaptateurs de langue L_S et L_T sont appris indépendamment en gelant les paramètres du Transformer.

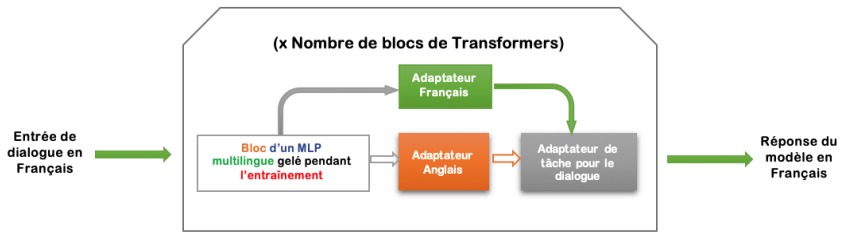

FIGURE 3 – Bloc du Transformer pour le **TrainOnSourceAdaptOnTarget**

3 Étude expérimentale

Afin de comparer et d'évaluer les trois approches présentées dans la section 2, nous avons affiné différent MLPs de L_S, L_T et une version multilingue.

Corpus d'entraînement : PersonaChat (Zhang *et al.*, 2018) est constitué d'un ensemble de dialogues entre deux humains en anglais, chacun se voyant attribué une personnalité initiale en quelques phrases.

Objectif d'entraînement : Nous avons utilisé un double objectif. D'une part la modélisation de langue avec comme entrée la concaténation de la personnalité, l'historique et la réponse. C'est uniquement sur cette dernière que la *fonction d'optimisation (loss function)* est calculée. D'autre part la classification multichoix, pour apprendre à sélectionner la bonne réponse dans un ensemble comprenant plusieurs distracteurs.

Description des modèles : Pour *TrainOnTarget*, nous avons affiné **GPT-fr** (Simoulin & Crabbé, 2021) (124M paramètres), comparable en termes d'architecture et de dimension au modèle Transfer-Transfo (Wolf *et al.*, 2019) basé sur **GPT-1** (117M) qui est utilisé pour *TestOnSource*. Enfin, pour *TrainOnSourceAdaptOnTarget*, le modèle utilisé est **BLOOM-560M**. Multilingue, il est le seul à permettre cette approche. Cette caractéristiques nous permet de l'utiliser également pour construire des modèles en L_S et en L_T pour les deux premières approches respectivement.

4 Évaluation

De part une structure *one-to-many* (Zhao *et al.*, 2017), l'évaluation automatique du dialogue n'est pas toujours en accord avec l'évaluation humaine, qui reste la plus fiable. Nous reportons ici les résultats de l'évaluation humaine accompagnés d'une analyse basée sur les conversations collectées.

Évaluation humaine : Nous avons collecté 120 conversations via l'interface RASA-X (Bocklisch *et al.*, 2017) en 2 phases : dans la 1ère, nous avons déployé `GPT-fr`, `GPT` et `BlenderBot 1` et dans la 2ième, les 4 modèles basés sur `BLOOM`. Elles ont ensuite été notées de 1 à 5 selon trois critères sélectionnés sur la base de ceux figurant dans (Mehri & Eskénazi, 2020; Ji *et al.*, 2022; Roller *et al.*, 2020) : la cohérence, l'engagement et le naturel. Les résultats sont reporté dans le tableau 1.

TABLE 1 – Différence de notes moyennes avec un modèle de référence (BlenderBot1) par critère

Stratégies	**Modèles**	**Cohérence**	**Engagement**	**Naturel**
Référence (TestOS)	**BlenderBot 1**	**3.64**	**4.45**	**3.77**
TrainOnSource	madx-BLOOM (L_S)	-2,10	-2,37	-2,06
AdapatOnTarget	madx-BLOOM (L_T)	-2,02	-2,14	-2,30
TestOnSource	BLOOM (L_S)	-1,72	-2,07	-1,82
	GPT	-1,59	-2,02	-2,10
TrainOnTarget	GPT-fr	-2,09	-2,49	-2,30
	BLOOM (L_T)	**-1,32**	***-2,01***	**-1,55**

`BlenderBot 1`, la référence est un modèle plus large ($\sim$ 2.7Mds paramètres, distillés en 400M)

appris avec des objectifs d'apprentissage complexes sur une grande variété de corpus. `BLOOM_fr` émerge comme le meilleur dans les trois catégories évaluées, en moyenne : **+0.26** pour la cohérence, **+0.01** pour l'engagement et **+0.56** pour le naturel par rapport à `GPT_EN`, son plus proche concurrent. Ce dernier a des notes proches de `BLOOM_en` avec un léger avantage sur la cohérence (+0.13) et l'engagement (+0.06) et un déficit sur le naturel (-0.29). Le dernier groupe est composé de `GPT_FR`, `madxBLOOM_fr`, `madxBLOOM_en` pour lesquels la note médiane pour tous les critères est proche de 1.5, i.e près de la moitié des conversations avec ces modèles ont reçu la note la plus basse possible.

Analyse des conversations : Le tableau 2 donne un autre aperçu des performances des modèles, en terme de nombre d'échanges. On observe la même tendance que précédemment : les modèles moins bien notés ont un plus faible nombre d'échanges par dialogue, puisque les utilisateurs avaient comme consignes de poursuivre au maximum la discussion. Cela peut expliquer aussi leurs notes d'engagement relativement faibles, ainsi que leurs scores de cohérence. En effet, suivant nos instructions, les conversations sont arrêtées dés que des comportements erratiques tels que des répétitions ou des hallucinations ont été observées par les utilisateurs.

TABLE 2 – Nombre de tours de parole moyen par dialogue

Modèles	BB1	GPT_FR	GPT	xBLOOM_fr	xBLOOM_en	BLOOM_fr	BLOOM_en
#échanges	**35,6**	15,4	**24,8**	12,9	20,6	**24,8**	**36,3**

BB1 = BlenderBot 1, xBLOOM = modèle avec architecture MAD-X et GPT = modèle anglais provenant de (Wolf *et al.*, 2019).

Nous avons utilisé le nombre de mots moyen par tour de parole pour évaluer le comportement relatif des testeurs vis-à-vis des modèles et inversement. Ces grandeurs ont été normalisées par testeur afin de s'affranchir des variabilités entre testeurs (personnalité, attente vis-a-vis d'un agent conversationnel (Walker *et al.*, 2021), habitudes,...) et ainsi observer des variations essentiellement liées aux modèles.

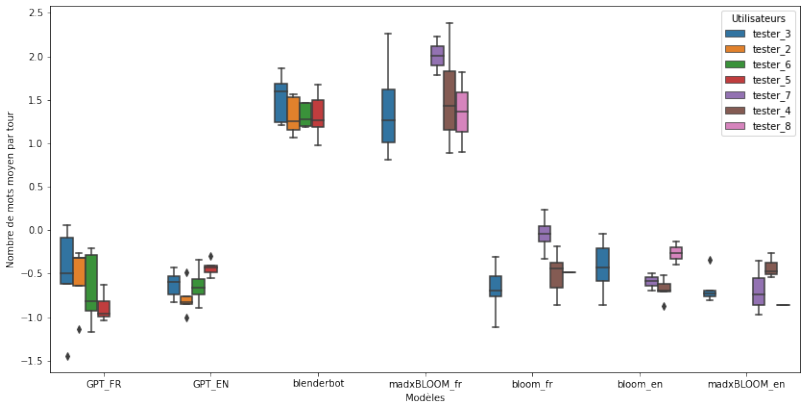

FIGURE 4 – **Nombre moyen de mots par tour de dialogue des differents modèles**

Dans la figure 4, on observe que `BlenderBot` se distingue de tous les autres à l'exception du modèle `madxBLOOM-fr` qui semble produire des réponses plus longue en moyenne. Ceci est dû au fait que, au moment du décodage, ce dernier a été contraint de générer au moins dix nouveaux tokens contrairement aux autres modèles. Enfin, globalement les différents modèles ont un comportement qui varie peu d'un utilisateur à autre. Cela met en évidence la problématique **P3** mentionnée dans (Bowden & Walker, 2023) à savoir un manque de personnalisation qui pourrait avoir un impact sur l'expérience utilisateur.

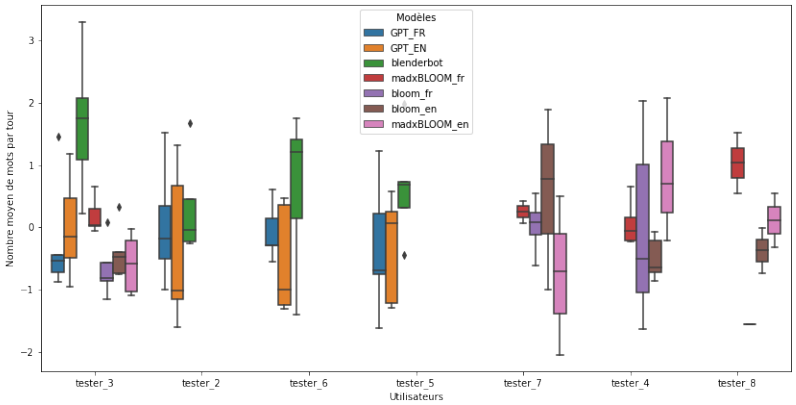

FIGURE 5 – **Nombre moyen de mots par tour de dialogue des testeurs**

Dans la figure 5 on observe *a contrario* que le comportement des utilisateurs varie beaucoup selon le modèle. Une analyse plus fine des dialogues concernés semble indiquer que pour `BlenderBot`, la référence, cela pourrait signifier un bon engagement de l'utilisateur tandis que pour `madxBLOOM-fr` (l'un des moins bien noté) cela pourrait illustrer les tentatives des utilisateurs de compenser la faible qualité des réponses en réorientant le modèle et en s'adaptant à lui. Aussi ce critère ne permet pas de distinguer clairement la qualité des modèles.

Prédiction des évaluations dans une configuration PARADISE (Walker *et al.*, 1997) : L'évaluation humaine implique un processus coûteux de collecte et annotations des conversations. Une solution est de prédire les notes d'évaluation directement à partir de mesures objectives sur les conversations. Pour vérifier cette possibilité, nous mettons en place une approche de type PARADISE. Le tableau 3 reporte les corrélations (r) entre les mesures objectives et les évaluations humaines.

On observe dans le tableau 3 que r est plus important avec les mesures directement dépendantes des modèles : **taille vocabulaire modèle** qui représente le nombre de lemmes différents utilisés par le modèle dans une conversation et le **nombre de mots moyens par tour de dialogue modèle**. Cependant, r avec les mesures indépendantes en plus d'être significative, restent très proches.

Nous avons entraîné différents modèles de régression des notations à partir des précédentes mesures normalisées (Régression linéaire, SVR, Arbre de Décision et MLP suivant ce qui a été fait dans (Walker *et al.*, 2021)). De notre jeu de 120 conversations, 10% sont aléatoirement assignées au jeu de test et nous entraînons les modèles sur celles restantes. Les notations sont elles aussi normalisées par évaluateur pour réduire le biais dans la notation. Dans le tableau 4 on observe des MSE relativement élevées sauf pour le critère engagement où on arrive avec le SVR à une MSE de **0.07**, un coefficient de détermination de **0.82** et surtout une corrélation de **92%** significative à $p \leq 0.001$. Avec les

TABLE 3 – Corrélations (r) entre les mesures objectives et les évaluations

Grandeurs mesurées	**Cohérence**	**Engagement**	**Naturel**
Nombre d'échanges	0.353	0.389	0.390
#Mots moyen par tour testeur	0.280	0.254	0.309
#Mots moyen par tour modèle	0.338	0.455	0.367
Taille vocabulaire testeur	0.429	0.457	0.468
Taille vocabulaire modèle	0.449	0.548	0.504
Taille vocabulaire conversation	0.466	0.526	0.510

Mesures significatives avec $p \leq 0,001$

mesures objectives utilisées, on ne pourrait donc prédire que les scores d'engagement avec une certaine qualité. Les évaluations humaines restent nécessaires, où d'autres mesures objectives, moins évidentes, doivent être mobilisées.

TABLE 4 – Résultats des modèles de régression sur les évaluations normalisées

	Cohérence			**Engagement**			**Naturel**		
Modèles	MSE	R^2	r	MSE	R^2	r	MSE	R^2	r
Regression linéaire	0.24	0.31	0.57	0.23	0.47	0.74*	0.27	0.41	0.65
SVR	0.34	0.05	0.48	**0.07**	**0.82**	**0.92****	0.35	0.22	0.57
Arbre de décision	0.42	-0.21	0.42	0.17	0.59	0.79*	0.41	0.096	0.43
MLP	0.28	0.20	0.55	0.21	0.53	0.74*	0.28	0.40	0.64

** significatif $p \leq 0.001$, * significatif $p \leq 0.01$

5 Conclusion

Le développement des modèles de dialogue à domaine ouvert en français est encore loin derrière les modèles anglais, ou même chinois, aujourd'hui. Il en va de même pour de nombreuses autres langues. La raison principale est le manque de corpus spécialisés. Cependant, la disponibilité de MLPs en français et d'outils TAN sont des atouts pouvant être mis à profit pour exploiter les ressources d'une langue plus dotée pour cette tâche. Dans cette optique, nous avons évalué trois approches différentes et comparé les modèles obtenus et un modèle de référence anglais utilisé avec un traducteur automatique. La stratégie TrainOnTarget avec un modèle multilingue a donné les meilleurs résultats (hors modèle de référence) lors de l'évaluation humaine. Ceci ouvre la voie à de futurs travaux sur l'utilisation de données traduites automatiquement avec des modèles multilingues tels BLOOM qui possèdent implicitement des capacités de traduction. L'amélioration des objectifs d'apprentissage pourrait alors permettre de rattraper les performances des modèles de référence des langues bien dotées pour la tâche, malgré l'obstacle que constitue la rareté des corpus spécifiques à chaque langue. Le fait qu'en dehors des dialogues à domaine ouvert, le français soit une langue bien dotée n'est pas totalement limitant pour ces approches. En effet, notre meilleur modèle était basé sur l'approche TrainOnTarget avec BLOOM, un modèle multilingue incluant plusieurs langues peu dotées et en accès libre.

Références

BIGSCIENCE WORKSHOP (2022). Bloom : A 176b-parameter open-access multilingual language model.

BOCKLISCH T., FAULKNER J., PAWLOWSKI N. & NICHOL A. (2017). Rasa : Open source language understanding and dialogue management. *CoRR*, **abs/1712.05181**.

BOWDEN K. K. & WALKER M. (2023). Let's get personal : Personal questions improve socialbot performance in the alexa prize.

JABAIAN B., BESACIER L. & LEFÈVRE F. (2013). Comparison and combination of lightly supervised approaches for language portability of a language understanding system. *IEEE Transactions on Audio, Speech and Language Processing*, **21**(3), 636–648.

JI T., GRAHAM Y., JONES G. J. F., LYU C. & LIU Q. (2022). Achieving reliable human assessment of open-domain dialogue systems. DOI : 10.48550/ARXIV.2203.05899.

LEFÈVRE F., MAIRESSE F. & YOUNG S. J. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In *Interspeech 2010, 11th Annual Conference of the International Speech Communication Association*, p. 78–81, Chiba, Japan.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). BART : denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

LIN Z., LIU Z., WINATA G. I., CAHYAWIJAYA S., MADOTTO A., BANG Y., ISHII E. & FUNG P. (2020). Xpersona : Evaluating multilingual personalized chatbot.

MEHRI S. & ESKÉNAZI M. (2020). Unsupervised evaluation of interactive dialog with DialoGPT.

PFEIFFER J., VULIĆ I., GUREVYCH I. & RUDER S. (2020). MAD-X : An adapter-based framework for multi-task cross-lingual transfer.

RADFORD A. & NARASIMHAN K. (2018). Improving language understanding by generative pre-training.

RASHKIN H., SMITH E. M., LI M. & BOUREAU Y.-L. (2019). Towards empathetic open-domain conversation models : a new benchmark and dataset. In *ACL*.

ROLLER S., DINAN E., GOYAL N., JU D., WILLIAMSON M., LIU Y., XU J., OTT M., SHUSTER K., SMITH E. M., BOUREAU Y.-L. & WESTON J. (2020). Recipes for building an open-domain chatbot.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle Transformer Génératif Pré-entrainé pour le _____ français. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Traitement Automatique des Langues Naturelles*, p. 246–255, Lille, France : ATALA. HAL : hal-03265900.

SMITH E. M., WILLIAMSON M., SHUSTER K., WESTON J. & BOUREAU Y. (2020). Can you put it all together : Evaluating conversational agents' ability to blend skills.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention Is All You Need.

WALKER M. A., HARMON C., GRAUPERA J., HARRISON D. & WHITTAKER S. (2021). Modeling performance in open-domain dialogue with PARADISE.

WALKER M. A., LITMAN D. J., KAMM C. A. & ABELLA A. (1997). PARADISE : A framework for evaluating spoken dialogue agents. *CoRR*, **cmp-lg/9704004**.

WOLF T., SANH V., CHAUMOND J. & DELANGUE C. (2019). Transfertransfo : A transfer learning approach for neural network based conversational agents.

ZHANG S., DINAN E., URBANEK J., SZLAM A., KIELA D. & WESTON J. (2018). Personalizing Dialogue Agents : I have a dog, do you have pets too ? *arXiv.org*.

ZHAO T., ZHAO R. & ESKENAZI M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. DOI : 10.48550/ARXIV.1703.10960.