Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

CQuAE : Un nouveau corpus de question-réponse pour l'enseignement

Thomas Gerald[1*] Louis Tamames[2*] Sofiane Ettayeb[2] Patrick Paroubek[1]
Anne Vilnat[1]
(1) Université Paris Saclay, CNRS, LISN
(2) Stellia
(1) prenom.nom@lisn.upsaclay.fr, (2) prenom.nom@stellia.ai

RÉSUMÉ

Dans cet article nous présentons un nouveau corpus de question-réponse en français pour le domaine de l'éducation. Ce corpus a été construit dans le but de créer un système d'assistant virtuel pour répondre à des questions sur des documents ou du matériel de cours. Afin d'être utile autant aux enseignants qu'aux étudiants, il est important de considérer des questions complexes ainsi que d'être capable de justifier les réponses sur du matériel validé. Nous présentons donc le nouveau corpus CQuAE, un corpus de questions-réponses manuellement annoté dont nous discutons des propriétés. Nous présenterons aussi les différentes étapes de sa création avec aujourd'hui une phase d'amélioration des données. Enfin, nous présentons plusieurs expériences pour évaluer l'exploitation du corpus dans le cadre d'un système de questions-réponses automatique. Ces différentes analyses et expériences nous permettrons de valider l'adéquation des données collectées pour l'objectif visé.

ABSTRACT

CQuAE : A new question-answering corpus for teaching assistant

In this article we present a new French question-answering corpus for the education domain. This corpus has been built with the aim of creating a virtual assistant system for answering questions on documents or course material. In order to provide a useful tool for both teachers and students, it is important to consider complex questions and to be able to justify answers based on validated material. We therefore present the new CQuAE Corpus, a manually annotated question-answering corpus and discuss its properties. We will also present the different stages in its creation, with a recent focus on data enhancement. Finally, we present experiments to evaluate the exploitation of the corpus within the framework of an automatic question-answering system. These different analyses and experiments will enable us to validate the suitability of the data collected for the intended purpose.

1 Introduction

Ce travail s'inscrit dans le domaine de l'éducation, notamment dans le but d'aider les étudiants à apprendre et à réviser leurs cours en leur fournissant des questions sur les documents de cours choisis par l'enseignant, ainsi que les réponses associées. L'objectif est de dépasser les simples questions factuelles auxquelles il est facile de répondre et de pouvoir poser des questions complexes qui vont au-delà de la recherche d'une réponse par entité nommée. Par exemple, face à un cours sur les débuts de la Révolution française, nous ne voulons pas seulement demander quand a eu lieu la prise de la Bastille, mais aussi quelles sont les raisons qui ont conduit les manifestants à le faire, nous rapprochant ainsi des questions de cours qu'un enseignant pourrait poser. Pour aider les enseignants dans cette

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 50–63.

démarche chronophage, nous avons travaillé à la constitution d'un tel corpus afin d'automatiser la tâche de création de questions et de réponses, en commençant par la construction manuelle d'un premier corpus.

À l'heure actuelle, aucun corpus ne répond à l'ensemble de ces critères, à savoir des questions et des réponses qui peuvent être complexes, qui s'appuient sur des ensembles de documents validés mais courts (les cours de l'enseignant), et qui sont en langue française. Nous souhaitons travailler sur plusieurs disciplines, éventuellement à différents niveaux d'enseignement, nous avons commencé notre étude par l'histoire telle qu'elle est enseignée à la fin du collège et au début du lycée. Pour avoir une base de comparaison, nous avons également effectué quelques tests sur la géographie, les sciences de la vie et l'éducation civique. Nous avons ainsi constitué un corpus contenant :

— des questions créées à partir de documents de cours, non seulement des questions factuelles mais aussi des questions plus complexes
— des réponses qui sont soit extraites du cours, soit construites à partir de plusieurs éléments disséminés dans le document,
— le document source, qui valide à la fois l'intérêt de la question et la qualité de la réponse produite.

Aujourd'hui nous en sommes à la seconde phase de constitution du corpus, nous avons déjà collecté plus de 11000 annotations que nous présentons dans cet article. Par ailleurs, nous décrirons aussi une phase de correction du corpus afin d'améliorer les données déjà collectées.

Avec ces différents éléments, ce corpus pourrait être utilisé pour former les composants d'un cadre de Génération Augmentée par Récupération (RAG). À cette fin, nous proposons dans ce travail de mesurer l'adéquation de l'ensemble de données pour développer une telle application via l'adaptation de modèles de langue.

Nous détaillerons d'abord comment nous avons collecté un nouveau corpus conçu pour cette tâche. Ensuite, nous présenterons une analyse de ce corpus afin d'en présenter le contenu. Puis, nous discuterons de la phase d'amélioration du corpus. Enfin, la partie suivante sera consacrée à la présentation des expériences que nous avons menées afin de démontrer la valeur de cet ensemble de données pour l'apprentissage d'un système de RAG, en comparant plusieurs grands modèles de langue et les différentes versions de notre corpus.

2 Travaux connexes

Les grands modèles de langue. La génération de résumés automatiques, de questions ou bien de réponses sont aujourd'hui des sujets centraux de recherche pour la communautés du TAL. Ces différentes tâches ont bénéficié des évolutions des algorithmes d'apprentissage profond. En particulier les architectures neuronales comme le modèle "Transformer" (Vaswani *et al.*, 2017) permettent aujourd'hui d'obtenir de bonnes performances pour ces différents objectifs.

Dans un premier temps les modèles de langues exploitant ces architectures ont été développés pour la langue anglaise, aujourd'hui plusieurs variantes en langue française existent comme les modéles *CammemBERT* ou *FlauBERT* (Martin *et al.*, 2020; Le *et al.*, 2020) pour la classification et le modèle *BARThez* pour des tâches de génération (Eddine *et al.*, 2021). Aujourd'hui la taille des modèles de langues permet de prendre en compte conjointement différentes langues, dans des approches multilingues (Scao *et al.*, 2022).

Adaptation des grands modèles de langue. Ces modèles de langue sont en général appris sur

des tâches de reconstruction de l'entrée ou de prédiction du prochain mot (ou jeton). Dès lors, il est nécessaire d'adapter ces modèles à la tâche visée. Pour adapter le modèles à une tâche, comme la classification ou la génération, une première méthode consiste à adapter les poids du réseaux de neurones, cette approche est connue sous le nom de *fine-tuning* (ou adaptation fine).

Cette approche est néanmoins sujette à différents inconvénients ; d'une part l'optimisation de l'intégralité des paramètres peut être prohibitive en temps de calculs ; d'autre part l'optimisation de tous les poids du modèles peut mener à des problèmes de sur-apprentissage.

Les méthodes appelées *adapter* tentent aujourd'hui de palier ces problèmes (Pfeiffer *et al.*, 2020) en introduisant dans les différents blocs de nouveaux sous-réseaux de neurones qui seront les seuls optimisés durant l'étape de *fine-tuning*. De plus ce type d'approche permet de conserver les poids du modèles original (appelé modèle pré-entraîné). Bien que le coût de l'adaptation soit diminué, le coût de l'inférence (l'étape de prédiction) est légèrement augmenté. Récemment une nouvelle approche connue sous le nom de *"LOw Rank Adaption"* (Hu *et al.*, 2022) propose à la fois de diminuer le coût d'adaptation tout en préservant un coût d'inférence similaire à celui du modèle pré-entraîné.

Les corpus de question-réponse. Chacune de ces approches pour adapter le modèle nécessite néanmoins des collections de données importantes. Nous allons présenter les différents corpus de question-réponse existants et leur intérêt pour la communauté. Le corpus SQuAD (Rajpurkar *et al.*, 2016) est l'un des premier corpus de grande taille pour l'apprentissage de modèle de question-réponse s'appuyant sur un contexte (la réponse devant s'y trouver) avec plus de 20.000 exemples. Plus récemment, Google a publié le corpus Natural Question (Kwiatkowski *et al.*, 2019) qui est un corpus de questions en langage naturel, avec des paragraphes longs et courts pour les réponses (extraits de la version anglaise de Wikipédia). Pour les approches de question-réponse conversationnelle, les corpus CANARD et QUAC (Elgohary *et al.*, 2019; Choi *et al.*, 2018) sont disponibles. Pour les questions-réponses basées sur la recherche de documents ou de passages, le corpus MSMarco (Nguyen *et al.*, 2016) est aujourd'hui une référence avec plus d'un million de questions. Si la plupart des corpus sont disponibles en anglais, la communauté française a également produit des corpus tels que FQuAD (d'Hoffschmidt *et al.*, 2020), Piaf (Keraron *et al.*, 2020) ou CALOR-QUEST (Bechet *et al.*, 2019) toujours dans le but d'extraire la réponse du contexte. Cependant, ces corpus reposent principalement sur des réponses factuelles, correspondant à un texte court, comme une entité nommée, un événement, une date, une quantité ou un lieu. Récemment, un nouveau corpus Autogestion (Antoine *et al.*, 2022) a été créé pour traiter les questions non factuelles, l'étude associée démontre l'incapacité des modèles standards à traiter les questions les plus complexes. Néanmoins à notre connaissance il n'existe aujourd'hui pas de corpus ouvert spécifique à l'enseignement secondaire pour des tâches de question-réponse complexes.

3 Le corpus

3.1 Récolte du corpus

Notre corpus se compose de paragraphes de cours, puis de questions et de réponses fondées sur ces textes. La première étape a donc été de recueillir un ensemble de textes en Français dans le domaine éducatif, en s'appuyant sur des livres scolaires (des cours) concernant les niveaux collège et lycée, principalement en Histoire mais aussi en Géographie, en Sciences de la Vie et de la Terre et Éducation Civique. Les premiers textes proviennent du site "le livre scolaire" [1]. Pour compléter

1. https ://www.lelivrescolaire.fr/

ce contenu, nous avons recherché des articles Wikipedia liés à ces sujets scolaires. Nous les avons filtrés en utilisant des API Wikipedia avec des requêtes construites à partir des titres de chapitres des livres, et en réunissant les sous-sections retournées. Pour ne pas avoir des contenus trop gros, nous avons découpé les articles en ne conservant pas plus de trois paragraphes par document. Un article Wikipedia va donc correspondre à plusieurs documents. Nous avons ainsi globalement réuni 3.891 documents (dont seulement 1.122 sont annotés), constitués de 14.433 paragraphes (dont seulement 3.893 sont annotés). Nous avons ensuite procédé à des campagnes d'annotation. Le principe est de présenter un paragraphe aux annotateurs en leur demandant de créer les annotations suivantes : (a) **une question** à poser ; (b) **le type de la question** qui peut-être factuelle, définition, cours ou synthèse ; (c) **le support de la question**, à savoir l'extrait du document à partir duquel la question est construite ; (d) **les éléments de réponse**, c'est à dire les passages permettant de répondre à la question ; (e) **la réponse** rédigée par l'annotateur, à partir des éléments précédents. Les annotateurs devaient créer environ 10 annotations (et plus si possible) pour chaque document. La Table 1 donne des exemples de questions, et de leurs supports.

L'un de nos objectifs principaux est de recueillir des questions requérant des niveaux d'expertise différents pour y répondre. Ce niveau de "difficulté" est lié au type de la question qui peut donc être :

- **factuelle** : la réponse est un fait ou une liste de faits (événement, personne, lieu, date...) ;
- **définition** : la réponse correspond à la définition d'un mot ou d'un concept ;
- **cours** : la réponse n'est pas réduite à un fait mais contient des explications ou des détails, qui doivent être explicites dans le document ;
- **synthèse** : la réponse s'appuie sur plusieurs éléments du document fournissant des informations diverses qui doivent être réunies ou impliquant une interprétation pour être produite.

Pour garantir d'avoir suffisamment de questions complexes, avec leurs réponses (autres que factuelles ou définitions), nous avons demandé un ratio de 40 % de factuelles et définitions, et 60 % questions de cours et de synthèses. Cependant, nous avons demandé également de ne pas créer artificiellement des questions complexes quand le document ne s'y prête pas, le ratio n'est donc pas strict. Les questions *factuelles* et *définition* sont assez simples à formuler et découlent directement du texte du document. Pour les questions de *cours*, elles sont un peu plus complexes et les réponses nécessitent plus de détails. La réponse aux questions de *synthèse* nécessite un raisonnement à partir du document.

Deux groupes d'annotateurs ont travaillé sur le corpus : le groupe A composé d'une vingtaine d'annotateurs personnes ayant un bon niveau général, mais sans expérience d'enseignement, le groupe B est composé de 6 annotateurs ayant une expérience en enseignement [2]. Dans la Table 2 nous indiquons la distribution actuelle en termes de type de questions pour les deux groupes. On notera que le groupe A a proportionnellement produit plus de questions de cours que le B, alors que le B s'est concentré sur les questions de synthèse. Pour aider les annotateurs, un guide a été fourni et est disponible avec le corpus sur le dépôt gitlab du corpus [3]. Aussi, nous reportons en annexe le nombre de questions et de documents par domaine et par source (*le livre scolaire* ou *wikipedia*).

3.2 Analyse du corpus

Pour une première analyse, nous avons comparé la longueur des questions et des réponses dans notre corpus, en fonction du type de la question, et en la comparant avec ce qu'on peut observer dans les deux corpus existant en français et déjà évoqués, à savoir FQuAD (d'Hoffschmidt *et al.*, 2020) et Piaf (Keraron *et al.*, 2020). La Figure 1 illustre cette étude.

2. Il est plus facile de recruter des personnes dans le groupe A que B, d'où la différence de taille des deux groupes.
3. https://gitlab.lisn.upsaclay.fr/gerald/cquae

Type	Question	Support
Factuelle	En quelle année Christophe Colomb a-t-il découvert l'Amérique ?	Christophe Columb découvre l' Amerique (1492)
Définition	Qu'est-ce qu'une presse rotative ?	Une presse rotative est une presse typographique montée sur un cylindre, permettant une impression en continu.
Cours	Comment les Européens ont-ils légitimé leur domination ?	Les Européens repensent la hiérarchie entre les peuples selon un schéma centré sur le christianisme et l'Europe qui leur sert ensuite à légitimer leur domination.
	Quels sont les noms de ceux qui indiquent comment pratiquer la religion musulmane ? Sur quel texte s'appuient-ils pour le faire ?	Ce sont les ulemas qui régissent la religion, en s'appuyant sur la loi de la Sharia.
Synthèse	Pourquoi certains français sont-ils favorables à l'état d'urgence après les attaques à Paris en 2015 ?	• il les protège de la menace terroriste et du risque d'une nouvelle attaque, qui est redoutée par tous. • Ce régime exceptionnel continue à apparaître comme une "nécessité"...
	Qui doit être impliqué pour lutter contre le changement climatique d'après Matt Petersen ? Comment ?	Matt Petersen travaille sur le développement soutenable dans la ville de Los Angeles, aux côtés du maire de la ville [...] nous avons besoin de tous. Tout sourire, le maire de Los Angeles a connecté [...] des panneaux solaires installé sur des toits privés [...] ...
	Pourquoi cet article parle de "victoire du féminisme" pour décrire le mouvement des midinettes ?	Il ne faut pas médire des midinettes. Il n'est pas d'un bon esprit de les taxer de frivolité parce qu'elles travaillent dans les robes, qu'elles sont jeunes et [...] de la femme, s'exerçant en ces jours tragiques au préjudice de milliers et de milliers d'ouvrières, d'employées, voire de fonctionnaires, est d'une si cruelle injustice qu'elle soulève des protestations de tous les côtés.

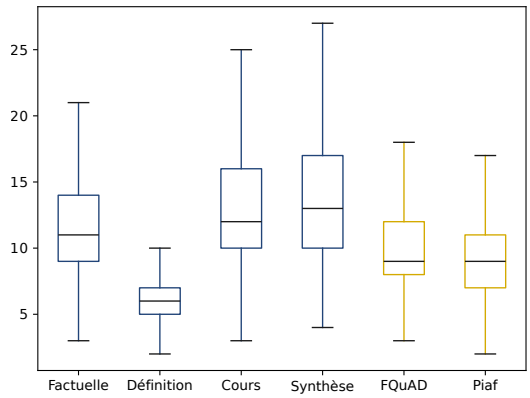
TABLE 1 – Exemples des quatre types de questions, avec leurs supports (extrait du document la justifiant)

Type de question	Groupe A	Groupe B	Total
Factuelle	2 106	294	2 400
Définition	1 506	181	1 687
Cours	4 784	490	5 274
Synthèse	1 756	338	2 094
Total	10 152	1 303	11 455

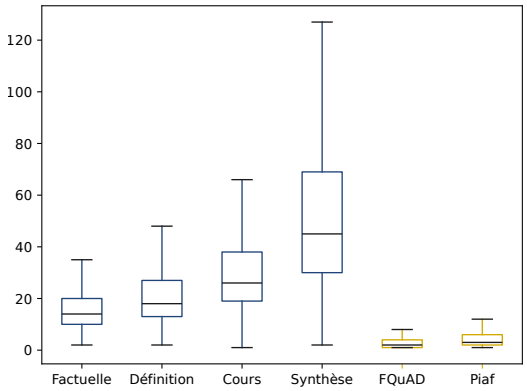
TABLE 2 – Statistiques pour les types de questions pour chaque groupe (A = éduqués, B = avec expérience de l'enseignement)

Les questions de *définition* sont toujours beaucoup plus courtes, les trois autres types sont assez comparables, les questions de synthèse (plus difficiles) étant un peu plus longues. Pour ces trois catégories l'écart-type est assez important. FQuAD et Piaf sont assez similaires, avec des questions plus courtes et moins d'écart-type. On note que les questions *factuelles* sont les plus proches de celles de ces deux datasets, ce qui confirme le fait que les questions y sont surtout factuelles. La même étude a été faite sur la longueur des réponses. Logiquement les réponses aux questions de *synthèse* sont plus longues que les autres, et parfois même très longues. Les réponses factuelles sont les plus courtes. Les réponses dans FQuAD et Piaf sont significativement plus courtes que les nôtres, même dans les *factuelles*. En effet, nous avons demandé une réponse rédigée aux annotateurs quant à FQuAD et Piaf la réponse est extraite du contexte. Notre corpus est pour cela différent des datasets existants.

Réponses extraites du contexte. Bien que nous demandions aux annotateurs de rédiger une réponse, nous avons remarqué qu'une grande partie des réponses sont directement issues du contexte. Nous nous proposons donc d'étudier la proportion de chaînes de caractères communes à la fois à la réponse et au contexte (paragraphes extraits sélectionnés par l'annotateur). Dans la table 3, nous reportons les résultats de cette analyse pour chacun des types de question. Nous capturons la plus grande chaîne de caractères commune entre la réponse et les paragraphes et nous en reportons la proportion par rapport à la taille de la réponse rédigée. Un coup d'oeil au résultats présentés dans le tableau nous montre que la proportion de mots communs est répartie différemment selon le type de la question. En



(a) Longueur de la question en fonction de son type dans notre corpus, comparée à FQuAD et Piaf



(b) Longueur de la réponse en fonction de son type dans notre corpus, comparée à FQuAD et Piaf

FIGURE 1 – Taille de la question ou de la réponse par type de question en comparaison du corpus Piaf et FQuAD

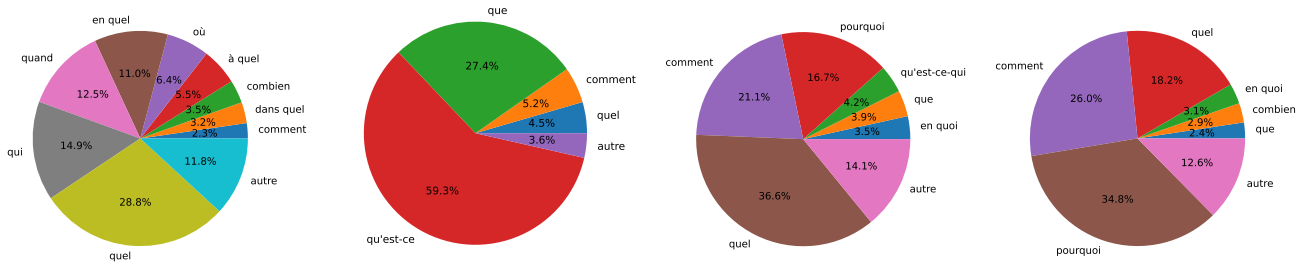
	Mediane	1^{er} quartile	3^{me} quartile
Factuelle	32.6 %	22.9 %	45.8 %
Définition	27.1 %	17.0 %	43.9 %
Cours	24.5 %	15.5 %	40.2 %
Synthèse	12.2 %	7.5 %	21.5 %
Total	24.4 %	14.6 %	39.5 %

TABLE 3 – Pourcentage de la plus longue chaîne de caractère commune à la réponse et au contexte.

particulier, les questions dites factuelles, comporte une forte proportion du contexte, cela s'explique par la réponse attendue ainsi que par la taille de la réponses.

Au contraire, les questions de synthèse comportent une proportion faible de contexte consécutif car les réponses nécessitent plusieurs éléments du texte et sont plus longues.

Caractérisation des types de questions. Pour caractériser les questions en fonction des différents types nous nous sommes demandés quels mots interrogatifs étaient utilisés dans la question. Pour extraire les mots interrogatifs, nous avons lémmatisé les questions et définis un certain nombre de formules interrogatives (quand, comment, où, etc...). Enfin nous avons regardé les correspondances avec les premiers lemmes de la question. Notons que pour 409 questions nous n'avons pas trouvé de correspondance, plusieurs raisons en sont la cause, d'une part certaines questions utilisent ces formes interrogatives mais pas au début de la question, par exemple :
"Selon Bartolomé de Las Casas, dans son livre Très brève relation de la destruction des Indes publié en 1552, quelle a été la raison du massacre des Indiens par les Espagnols lors de la conquête des Indes sous le règne de Charles Quint ?"
Par ailleurs, certaines structures ne sont pas capturées ou ne sont pas des questions comme par exemple :
"En vous basant sur ces documents, citez un argument en faveur de la loi et un autre contre."
Nous reportons les résultats des différentes formes interrogatives capturées pour chaque types de question dans la figure 2. On remarquera que les distributions des formes interrogatives sont dépendantes du type de la question ; ainsi pour les questions de synthèse les mots clefs *"pourquoi"* et *"comment"* sont prépondérants, ceux-ci impliquant une explication dans la réponse ; pour les questions de définition on retrouvera le mot *qu'est-ce* menant vers une description d'un concept. Pour les



(a) Factuelle (b) Définition (c) Cours (d) Synthèse

FIGURE 2 – Comparaison des mots de la question pour les différents types de question

questions factuelles les mots clefs comme *"quel"*, *"qui"*, *"quand"* ou *"en quel"* entraînent une réponse de type entité nommée. Pour les questions de cours en revanche nous avons une panoplie de mots interrogatifs amenant à une réponse explicative ou factuelle. Cette toute première étude sera enrichie par la suite.

4 Correction et amélioration des annotations

Suite à une première pré-évaluation du corpus, nous avions remarqué que certaines questions et/ou réponses contenaient des erreurs ou des imprécisions. Nous avons remarqué plusieurs types d'erreurs dans les données : des erreurs de syntaxe ; des erreurs liées à l'ajout d'informations non présentes dans le document ou à l'inverse des réponses incomplètes ; des questions en dehors du sujet ou non pertinentes dans le cadre de l'éducation. Par ailleurs nous avons émis l'hypothèse que de telles erreurs pourraient être à l'origine d'imprécisions dans l'évaluation et l'entraînement de systèmes de questions et réponses automatique. Pour répondre à cette problématique, nous nous sommes proposés de mettre en place une seconde campagne, visant à évaluer le taux d'annotations comportant des erreurs.

Évaluation des données erronées. Pour évaluer la pertinence des données annotées, cinq annotateurs ont été mis à contribution pour juger un ensemble de 9243 annotations. Nous leur avons proposé d'annoter les questions et réponses en suivant les critères binaires suivants :

- **Q+ :** la question est correcte
- **Q- :** la question est corrigeable (syntaxe, reformulation possible)
- **R+ :** la réponse est correcte
- **R- :** la réponse est corrigeable (syntaxe, mauvaise entités, etc...)
- **HS :** le couple question/réponse est non pertinent et ne peut pas être corrigé

Pour cette évaluation nous reportons les résultats dans le table 4 conditionnellement au type de question. Pour les erreurs dans les questions, il semblerait que celles de type définitions comportent moins d'erreur, avec plus de 81 % d'annotations ne nécessitant aucune correction. Pour les autres catégories de questions, les erreurs semblent être réparties de manière similaire (entre 65 % et 68 %). Pour les erreurs dans les réponses, ce sont les catégories cours et synthèse qui nécessitent le plus grand nombre de corrections ; les réponses étant plus longues, des fautes de syntaxe sont plus probables.

Sur les 9243 annotations 192 ont été évaluées par tous les annotateurs, nous permettant de calculer l'accord inter-annotateur. Nous avons obtenu un kappa de Fleiss [4] pour le critère "la question est-elle

4. https://fr.wikipedia.org/wiki/Kappa_de_Fleiss

	Q+	Q-	R+	R-	HS
Factuelle	65.4	28.5	62.0	31.8	5.9
Définition	81.1	14.6	64.9	30.4	4.2
Cours	67.7	26.6	52.6	41.4	5.5
Synthèse	65.7	28.1	49.4	43.5	6.3
Tout type	68.9	25.4	55.8	38.1	5.6

TABLE 4 – Pourcentages de questions (Q) et de réponses (R) correctes (+) ou corrigeables (-) et de couples Q-R non pertinents (HS) par type de question lors de l'évaluation de notre corpus

correcte ?" de .30 (avec un accord maximum kappa de Cohen [5] de .44 et minimum de .18), pour la réponse on obtient .18 (accord maximum de .40 et minimum de .29). Il est clair que nous n'avons pas un accord probant entre les annotateurs, ces résultats démontrent ainsi la difficulté de la tâche. Pour se faire une idée de la pertinence de l'évaluation des documents corrigés, nous proposerons dans la section suivante une évaluation des corrections afin de juger de la pertinence des modifications apportées.

Correction du corpus. Dans un second temps, nous nous sommes proposés de corriger les annotations ayant au moins la réponse ou la question évaluée comme corrigeable. Nous avons pu à ce jour revoir 2565 réponses et 1840 questions sur 3140 annotations. En définitive, nous avons un corpus composé de 8180 questions et réponses vérifiées.

Bien que pour la phase d'évaluation, il est difficile de parler d'un réel accord entre les annotateurs, nous pensons que les corrections effectuées tendent bien vers l'amélioration du corpus. Pour vérifier cette hypothèse, nous nous sommes proposés d'évaluer les corrections manuellement. Pour ce faire nous avons demandé à trois personnes de déterminer d'après eux, quel est le meilleur couple de question et réponse. Notons que dans certains cas aucune préférence n'a été exprimée, nous considérons cette possibilité comme un troisième label. Nous avons obtenu pour cette évaluation avec 3 annotateurs une préférence pour les éléments corrigés dans 76. % des cas, pour les annotations originales 15.3 % étaient préférées et dans 8.3 % des cas il n'y avait aucune préférence. Pour ces différents résultats nous obtenons un accord inter-annotateurs moyen de .51. Notons que si nous supprimons les cas où les annotateurs sont restés indécis nous obtenons un accord de .62 montrant ainsi un accord important. Ces résultats montrent donc l'intérêt d'avoir corrigé certaines annotations pour améliorer le corpus.

5 Génération de réponses

Le corpus annoté a pour objectif la création d'un assistant scolaire pour les thématiques d'enseignement secondaire, particulièrement sous la forme d'un système de génération de questions et/ou réponses. Ainsi il nous parait pertinent de juger de la qualité de la génération de ces modèles pour nos données. Dans cette section nous nous proposons d'évaluer automatiquement la tâche de génération de réponse étant donnée un document et une question. Notons aussi, que dans le cas d'un assistant pour l'enseignement, la recherche de documents à partir d'un corpus vérifié (mis à disposition par l'enseignant) est une des tâches nécessaire au fonctionnement du système. Ainsi, nous proposerons de comparer la dernière configuration considérant les documents cibles donnés par les annotateurs ou retrouvés via la méthode BM25(Robertson & Zaragoza, 2009).

5. https://fr.wikipedia.org/wiki/Kappa_de_Cohen

Qtype	Entrainement-v1	Entrainement-v2	Validation-v2	Test-v2
Factuelle	2144	1409	128	128
Définition	1431	1075	128	128
Cours	5018	3409	128	128
Synthèse	1838	1263	128	128
Tous	10431	7156	512	512

TABLE 5 – Nombre de couples Q-R utilisés pour l'adaptation des modèles de langue provenant de la version corrigée (v2) ou non (v1), en fonction du type de question

5.1 Protocole expérimental

Modèles. Nous nous sommes penchés sur l'entraînement de grand modèle de langue de 7 milliards de poids, choisissant ainsi Llama2 7b et Mistral 7b. Dans l'ensemble des entraînements, nous avons utilisé LoRA et la quantification des modèles en 8 bits (paramètres en annexe), cela permet de réduire le temps de calcul et d'inférence avec une perte en performance minime. Nous avons fixé la longueur maximale des phrases à 2048 tokens car l'entrée et la sortie maximale en utilisant nos données ne dépasse pas les 2000 tokens. Les modèles ont été entraînés sur 3 *epochs* sur l'ensemble du dataset, et nous avons sélectionné la sauvegarde avec la meilleure *loss* pour l'évaluation.

Ensemble d'entraînement et de test. Les ensembles d'entraînement utilisés sont des sous ensembles de la première version du corpus (v1) et de la version corrigée (v2). De plus, les ensembles de validation et de test sont extraits de la v2, respectivement pour déterminer l'arrêt de l'apprentissage ainsi que pour calculer les résultats présentés dans ce document. Nous obtenons ainsi les différents ensembles reportés dans le tableau 5.

Métriques pour l'évaluation. Pour évaluer la capacité des modèles à répondre aux questions, nous avons utilisé trois métriques principales : ROUGE, BERTScore, et GPT. ROUGE mesure la similarité entre les réponses générées et les réponses de référence en évaluant la correspondance des n-grams, utile pour apprécier la proximité des générations au contenu. BERTScore compare la similarité sémantique à l'aide des plongements de mots. L'évaluation par GPT (version gpt-3.5-turbo-0125) se fait en donnant une réponse de référence et une générée pour obtenir une note de 0 à 10 mais aussi une explication sur la qualité suivant des critères que nous avons définis. Ces métriques sont loin d'être parfaites, mais permettent de nous donner une intuition sur la pertinence des réponses générées.

5.2 Génération de réponse en contexte

	Paragraphes de la question (gold)				Paragraphes retrouvés par BM25			
Model	GPT3.5	B-Score	R-1	R-L	GPT3.5	B-Score	R-1	R-L
MISTRAL	**8.45**	0.789	0.399	0.345	**7.82**	0.762	0.342	0.29
MISTRAL-V1	8.40	**0.854**	**0.605**	**0.54**	7.62	**0.826**	**0.518**	**0.459**
MISTRAL-V2	8.41	0.852	0.598	0.532	7.62	0.825	0.51	0.451
LlaMA	**8.53**	0.717	0.183	0.148	**8.05**	0.699	0.153	0.124
LlaMA-V1	8.37	**0.848**	**0.587**	0.527	7.58	**0.822**	**0.509**	**0.453**
LlaMA-V2	8.33	0.846	0.584	0.52	7.63	0.821	0.504	0.447

TABLE 6 – Résultats pour la génération de réponses étant donné un contexte pris dans le corpus d'origine, ou retrouvé par BM25

Dans cette expérience nous tentons d'évaluer la capacité des modèles à produire des questions et

réponses en utilisant les informations contenues dans les paragraphes du corpus. Pour ce faire, nous proposons d'adapter les deux modèles Llama et Mistral en utilisant avec les configurations données dans la section 5.1. Nous utilisons les modèles *chat* et *instruct* de Meta et Mistral, qui sont faits pour répondre à une instruction. Nous utilisons pour l'entraînement et l'évaluation le prompt suivant :

"En se basant exclusivement sur le document, répondez à la question. La réponse est destinée à un élève voulant améliorer sa compréhension du cours. : {question} document :{documents} Réponse :".

Nous comparons aussi les résultats des modèles appris sur la v1, la v2 et le modèle pré-entraîné. Les différents résultats sont reportés dans le tableau 6.

Exceptée la métrique basée sur GPT, les résultats sont significativement meilleurs en adaptant les modèles sur notre corpus [6], ce phénomène indiquant que l'adaptation est nécessaire pour espérer avoir des réponses semblables à celles de notre corpus. Notons aussi que l'ensemble d'entraînement de la version corrigée est amputé d'environ 30 % des exemples, malgré cela, les performances entre les deux versions sont comparables [7]. Si l'on s'intéresse à la métrique se basant sur le score donné par GPT, l'interprétation est ici différente. En effet, les scores attribués au modèle adapté avec vérité terrain (première partie du tableau) sont similaires à ceux produits par le modèle pré-entraîné. Lorsque les paragraphes correspondent à ceux données par l'approche BM25 (un seul paragraphe), ce score est alors favorable au modèle non adapté. Bien que nous puissions difficilement en affirmer la cause, plusieurs hypothèses sont possibles d'après nos observations. D'une part, le modèle non-adapté produit souvent des réponses en langue anglaise, le score donné par le modèle GPT ne pénalise pas la langue utilisée. Une deuxième hypothèse est que le modèle non-adapté pourrait produire des réponses dont les éléments ne sont pas présents dans le contexte. Cette dernière hypothèse est appuyée par la différence de score des modèles utilisant les deux configurations des paragraphes.

Nous savons que les documents retrouvés ne correspondent pas intégralement à des documents pertinents d'après les résultats de la méthode BM25 (62 % des paragraphes étant dans la vérité terrain). En incluant BM25, les performances supérieures pour le modèle non-adapté laissent supposer que ce modèle est moins dépendant du contexte.

Notons enfin que la métrique GPT est difficile à exploiter, celle-ci dépendant à la fois du prompt, du modèle GPT et de la graine de génération. Il est donc difficile avec les éléments dont nous disposons de statuer sur les hypothèses proposées.

6 Conclusion

Dans cet article nous avons présenté un nouveau corpus CQuAE de question-réponse dans le domaine de l'enseignement secondaire. Après avoir analysé les versions du corpus recueilli, nous avons constaté les difficultés qu'il y avait à en garantir la qualité, du fait de la difficulté de la tâche. Nous avons discuté des améliorations de ce corpus avec une phase de correction des annotations. Enfin nous avons proposé d'étudier la pertinence de ces modèles pour la génération de réponses avec des grands modèles de langue. Avec les différentes analyses et conclusions sur les évaluations humaines nous pensons qu'un tel corpus serait bénéfique pour la communauté enseignante et celle du TAL. Cependant, avec les résultats dont nous disposons, il reste difficile de statuer sur les différentes hypothèses émises. Nous planifions donc d'explorer les différents résultats obtenus via une évaluation humaine. Actuellement, la version corrigée du corpus comporte un nombre d'exemples réduit par

6. $pvalue$ pour T-test d'indépendance < 0.01
7. $pvalue$ pour T-test d'indépendance > 0.05

rapport au corpus originalement produit. Pour ce dernier point, nous avons prévu de réviser les exemples manquant dans le corpus corrigé.

Références

ANTOINE E., AUGUSTE J., BÉCHET F. & DAMNATI G. (2022). Génération de questions à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents. *29e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.

BECHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019). CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019*.

CHOI E., HE H., IYYER M., YATSKAR M., YIH W., CHOI Y., LIANG P. & ZETTLEMOYER L. (2018). Quac : Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics.

D'HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). Fquad : French question answering dataset. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 de *Findings of ACL*, p. 1193–1208 : Association for Computational Linguistics. DOI : 10.18653/V1/2020.FINDINGS-EMNLP.107.

EDDINE M. K., TIXIER A. J. & VAZIRGIANNIS M. (2021). Barthez : a skilled pretrained french sequence-to-sequence model. In *EMNLP (1)*.

ELGOHARY A., PESKOV D. & BOYD-GRABER J. L. (2019). Can you unpack that ? learning to rewrite questions-in-context. In *EMNLP-IJCNLP* : Association for Computational Linguistics.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). Lora : Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* : OpenReview.net.

KERARON R., LANCRENON G., BRAS M., ALLARY F., MOYSE G., SCIALOM T., SORIANO-MORALES E. & STAIANO J. (2020). Project PIAF : building a native french question-answering dataset. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, p. 5481–5490 : European Language Resources Association.

KWIATKOWSKI T., PALOMAKI J., REDFIELD O., COLLINS M., PARIKH A. P., ALBERTI C., EPSTEIN D., POLOSUKHIN I., DEVLIN J., LEE K., TOUTANOVA K., JONES L., KELCEY M., CHANG M., DAI A. M., USZKOREIT J., LE Q. & PETROV S. (2019). Natural questions : a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, **7**, 452–466. DOI : 10.1162/TACL_A_00276.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : Unsupervised language model pre-training for french. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of The 12th Language Resources and Evaluation Conference,*

LREC 2020, Marseille, France, May 11-16, 2020, p. 2479–2490 : European Language Resources Association.

MARTIN L., MÜLLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. R. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 7203–7219 : Association for Computational Linguistics. DOI : 10.18653/V1/2020.ACL-MAIN.645.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). MS MARCO : A human generated machine reading comprehension dataset. In T. R. BESOLD, A. BORDES, A. S. D'AVILA GARCEZ & G. WAYNE, Éds., *Proceedings of the Workshop on Cognitive Computation : Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 de *CEUR Workshop Proceedings* : CEUR-WS.org.

PFEIFFER J., RÜCKLÉ A., POTH C., KAMATH A., VULIĆ I., RUDER S., CHO K. & GUREVYCH I. (2020). AdapterHub : A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 46–54, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.7.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100, 000+ questions for machine comprehension of text. In J. SU, X. CARRERAS & K. DUH, Éds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, p. 2383–2392 : The Association for Computational Linguistics. DOI : 10.18653/V1/D16-1264.

ROBERTSON S. & ZARAGOZA H. (2009). The probabilistic relevance framework : Bm25 and beyond. *Found. Trends Inf. Retr.*, **3**(4), 333–389. DOI : 10.1561/1500000019.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIC S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., GOKASLAN A., SIMHI A., SOROA A., AJI A. F., ALFASSY A., ROGERS A., NITZAV A. K., XU C., MOU C., EMEZUE C., KLAMM C., LEONG C., VAN STRIEN D., ADELANI D. I. & ET AL. (2022). BLOOM : A 176b-parameter open-access multilingual language model. *CoRR*, **abs/2211.05100**. DOI : 10.48550/ARXIV.2211.05100.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. VON LUXBURG, S. BENGIO, H. M. WALLACH, R. FERGUS, S. V. N. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 5998–6008.

7 Annexe

7.1 Sources et domaines

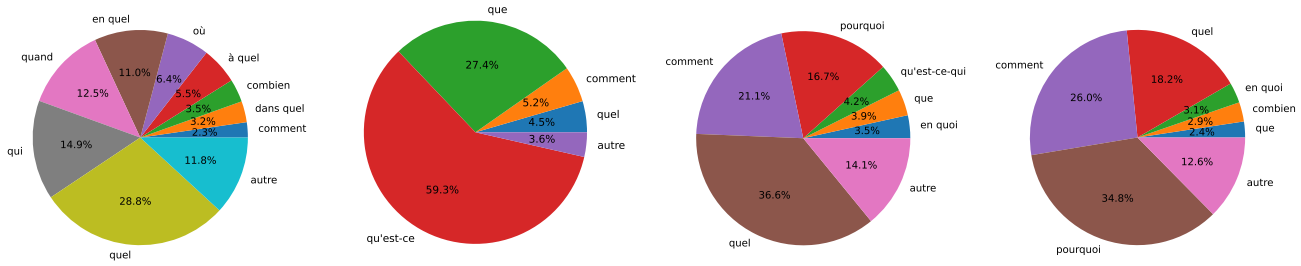
Dans la table 7 nous reportons le nombre de questions et de documents en fonction de la source, “lelivrescolaire” (préfixé par `lls`) et “wikipedia” (préfixé par `wik`), du domaine (géographie, histoire, svt) et du niveau .

Source	domaine-niveau	Version	nombre de questions	nombre de documents
lls	geographie-premiere	V1	572	48
		V2	409	48
lls	geographie-seconde	V1	716	64
		V2	533	64
lls	histoire-geographie-cinquieme	V1	816	86
		V2	588	86
lls	histoire-geographie-sixieme	V1	814	89
		V2	599	89
lls	histoire-premiere	V1	945	101
		V2	687	101
lls	histoire-seconde	V1	1146	103
		V2	782	102
lls	svt-cinquieme	V1	107	9
		V2	78	9
lls	svt-seconde	V1	204	18
		V2	148	18
wik	geographie-premiere	V1	1786	193
		V2	1260	189
wik	geographie-seconde	V1	546	51
		V2	374	50
wik	histoire-premiere	V1	1160	132
		V2	849	130
wik	histoire-seconde	V1	1996	155
		V2	1390	155
wik	histoire-geographie-cinquieme	V1	170	15
		V2	118	14
wik	histoire-geographie-sixieme	V1	184	17
		V2	139	17
wik	svt-cinquieme	V1	158	18
		V2	118	18
wik	svt-seconde	V1	135	23
		V2	108	23

TABLE 7 – Nombre de questions et documents par source et par domaine

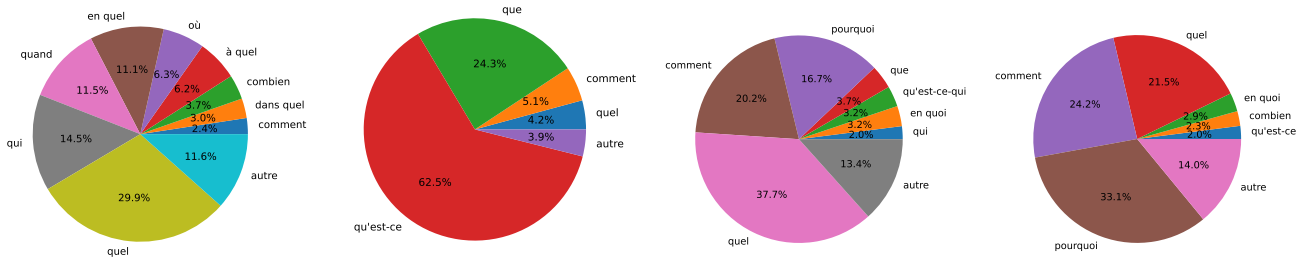
7.2 Comparaison des distributions du corpus avec FQuAD et PIAF

Les figures 3 et 4 décrivent les distributions des mots de la question pour les deux versions du corpus. Dans la figure 5 nous observons la distribution des mots de la question pour les corpus FQuAD (d’Hoffschmidt *et al.*, 2020) et Piaf (Keraron *et al.*, 2020). On pourra remarquer la similitude entre les distribution Piaf et FQuAD avec les questions “factuelles”. Sur les autres types, définition, cours, synthèse les distributions sont éloignées.



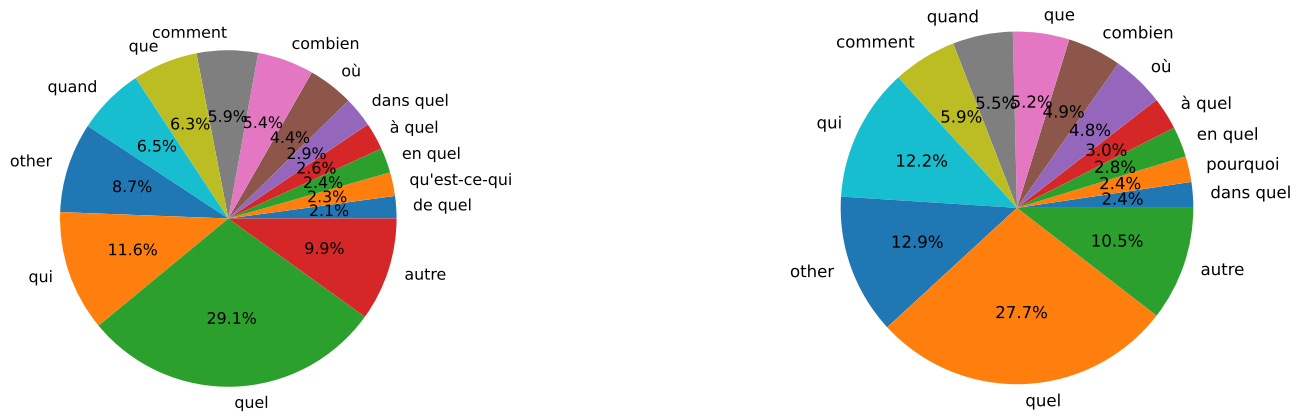
(a) Factuelle (b) Définition (c) Cours (d) Synthèse

FIGURE 3 – Comparaison des mots de la question pour les différents types de question sur la première version du corpus (voir section 3.2



(a) Factuelle (b) Définition (c) Cours (d) Synthèse

FIGURE 4 – Comparaison des mots de la question pour les différents types de question sur la seconde version du corpus (voir section 3.2



(a) FQuAD (b) Piaf

FIGURE 5 – Mots de la question pour les corpus FQuAD et Piaf