Ce document a été mis en ligne par l'organisme FormaV®.

Toute reproduction, représentation ou diffusion, même partielle, sans autorisation préalable, est strictement interdite.

Pour en savoir plus sur nos formations disponibles, veuillez visiter :
formav.co/explorer

LICENCE

SCIENCES DU LANGAGE

C9 : MENER UN TRAVAIL SUR CORPUS DE DONNÉES LANGAGIÈRES

SESSION 2026

Durée : 3 heures

SPÉCIALITÉ / UNITÉ	COEFF
SCIENCES DU LANGAGE	3,0 (COEFFICIENT INDICATIF)

Moyens de calculs et matériel autorisés :

L'usage de calculatrice avec mode examen actif est autorisé. L'usage de calculatrice sans mémoire, « type collège », est autorisé.

Aucun document ou matériel autorisé en dehors de ceux fournis avec le sujet. Un ordinateur n'est pas nécessaire.

Dès que le sujet vous est remis, assurez-vous qu'il est complet.

Ce sujet comporte 11 pages numérotées de 1/11 à 11/11.

Les documents réponses en fin de sujet sont à rendre avec la copie.

DOSSIER SUJET

MISE EN SITUATION ET CONSIGNES GÉNÉRALES DE L'ÉVALUATION : CELLULE CORPUS DE L'UNIVERSITÉ DES RIVES, LYON, SITE D'ÉTUDE DÉFINI POUR L'ÉPREUVE

La Cellule Corpus étudie les marqueurs discursifs employés dans des échanges de demande d'information. Vous devez documenter et analyser l'emploi de *du coup* et de *donc* dans deux sous-corpus anonymisés, puis proposer un protocole reproductible de traitement des données.

Le corpus et les structures décrits dans le dossier sont des données de cas fournies pour cette épreuve. Les modalités d'évaluation d'une licence Sciences du langage sont déterminées par l'établissement qui la délivre. Le code C9, la durée de 2 heures et le coefficient indicatif 3,0 inscrits ici ne constituent pas un code national d'épreuve ni des modalités nationales communes aux universités.

Répondez dans l'ordre des parties. Justifiez les calculs, indiquez les unités et distinguez les observations descriptives des interprétations. Les réponses aux tableaux des documents DR 1 et DR 2 sont à rendre avec la copie.

PARTIE 1 : Définition de l'objet et cadre de collecte (5 points)

Analyse fonctionnelle, constitution du corpus et cadre réglementaire

Vous intervenez sur un corpus constitué à partir d'échanges enregistrés dans un service d'accueil. Les deux sous-corpus sont désignés A et B. Les participants ont été informés de la collecte et ont donné leur accord. Les fichiers fournis ont été pseudonymisés; aucune identité directe n'est communiquée dans le sujet. Les paramètres de collecte et les extraits figurent dans le DT 1 et le DT 2.

Données de cadrage des deux sous-corpus

Sous-corpus	Situation d'interaction	Nombre d'échanges	Durée audio	Nombre de tokens transcrits
A	Demande d'information au guichet	24	12 min	7 200
B	Demande d'information par téléphone	36	18 min	10 800

Données agrégées fournies par la Cellule Corpus pour l'analyse.

1.1 (2 points) Formulez une problématique comparative portant sur l'emploi de *du coup* et de *donc*. Précisez une hypothèse vérifiable et deux variables à contrôler ou à discuter pour comparer les sous-corpus A et B.

1.2 (1,5 point) À partir du DT 1, expliquez pourquoi les deux sous-corpus ne doivent pas être comparés uniquement à partir de leurs nombres bruts d'occurrences. Indiquez la mesure de fréquence adaptée et son unité.

1.3 (1,5 point) À partir du DT 2, indiquez deux précautions concrètes nécessaires avant de diffuser un extrait ou un fichier du corpus. Pour chacune, reliez la précaution à un risque identifié.

PARTIE 2 : Exploitation quantitative et interprétation (6 points)

Calculs de fréquences, contrôle des données et interprétation linguistique

Les occurrences validées après vérification manuelle sont indiquées dans le DT 1. Pour une fréquence normalisée, utilisez la formule $F = \frac{n}{N} \times 1000$, où n est le nombre d'occurrences du marqueur et N le nombre de tokens du sous-corpus. Exprimez le résultat en occurrences pour 1 000 tokens et arrondissez au centième.

Occurrences validées des marqueurs

Sous-corpus	Tokens	Occurrences de « du coup »	Occurrences de « donc »
A	7 200	42	18
B	10 800	36	54

Les comptes portent sur les emplois discursifs validés, hors emplois littéraux.

2.1 (2 points) Calculez la fréquence normalisée de *du coup* dans A et dans B, puis celle de *donc* dans A et dans B. Présentez les quatre calculs avec leur unité.

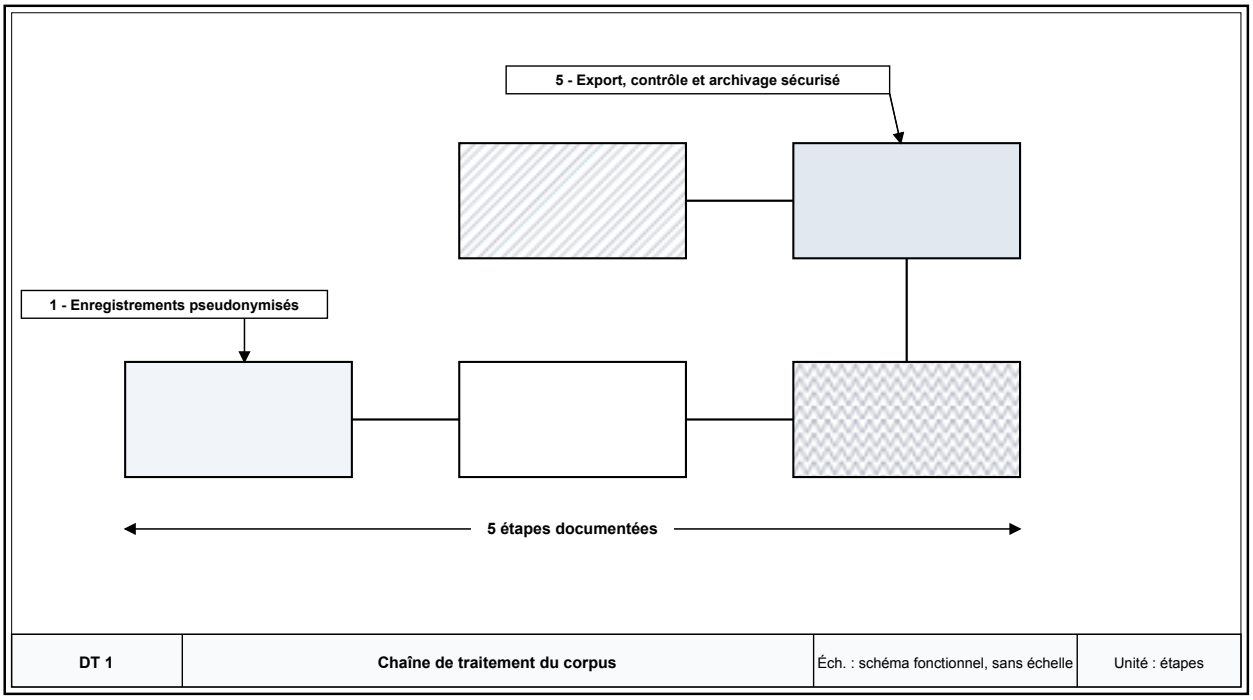
2.2 (2 points) Calculez, pour chaque marqueur, le nombre total d'occurrences et la fréquence globale dans les 18 000 tokens réunis. Complétez le DR 1. Détaillez au moins un calcul de fréquence globale.

2.3 (2 points) Comparez les fréquences observées dans A et B. Donnez une interprétation prudente des écarts et une limite de l'inférence possible à partir des seuls nombres fournis. Appuyez votre réponse sur au moins deux résultats numériques.

PARTIE 3 : Préparation d'un protocole de traitement reproductible (5 points)

Chaîne de traitement, annotation et contrôle de qualité

La Cellule Corpus doit préparer une version analysable des enregistrements et documenter les décisions d'annotation. Le DT 1 décrit les étapes prévues; le DT 3 fournit les conventions et les données de contrôle disponibles. Les durées de traitement données dans le DT 3 sont des estimations pour les fichiers de ce cas.



3.1 (2 points) Décrivez dans l'ordre les opérations nécessaires entre la réception des enregistrements et la production d'un fichier exploitable pour l'analyse. Identifiez au moins quatre opérations et indiquez pour chacune un contrôle de qualité ou une trace à conserver.

3.2 (2 points) Le DT 3 prévoit 12 minutes d'audio pour A, 18 minutes pour B et un temps moyen de transcription de 4,5 minutes de travail par minute d'audio. Calculez le temps total de transcription en minutes et en heures. Au tarif de 24 euros par heure, calculez le coût de transcription, hors charges et relecture. Présentez les formules $T = D \times r$ et $C = T_h \times c_h$ en appliquant les unités adaptées.

3.3 (1 point) Le DT 3 indique 172 décisions concordantes sur 200 unités annotées en double. Calculez le taux d'accord observé. Expliquez en une phrase pourquoi ce résultat ne suffit pas, à lui seul, à établir la fiabilité complète de l'annotation.

PARTIE 4 : Prévention, sécurité et responsabilité dans le travail sur corpus (4 points)

Protection des personnes, sécurité des données et prévention des risques de travail

Le traitement du corpus implique l'accès à des données issues d'interactions humaines, l'usage d'un poste informatique et la conservation de fichiers audio et de transcriptions. Les consignes de sécurité et de protection des données sont rappelées dans le DT 2 et le DT 3.

Risques à prendre en compte dans le traitement du corpus

Situation	Risque possible	Mesure à proposer
Copie d'un fichier audio sur un espace partagé non contrôlé	Accès non autorisé ou réidentification	.....
Séance longue de transcription sans interruption	Fatigue visuelle et inconfort postural	.....
Diffusion d'un extrait contenant un nom propre	Divulgation d'une donnée identifiante	.....

Associer chaque risque à une mesure de prévention adaptée.

4.1 (2 points) Pour les trois situations du tableau de la partie 4, proposez une mesure de prévention concrète et proportionnée. Complétez le DR 2. Les mesures doivent distinguer la sécurité des fichiers, la protection des personnes et la prévention des risques liés au poste de travail.

4.2 (1 point) Citez deux informations qu'il convient de vérifier dans le dossier de collecte avant toute réutilisation ou diffusion des données. Expliquez brièvement l'importance de chacune.

4.3 (1 point) Indiquez une règle de nommage ou de versionnage qui permet de retrouver les fichiers de travail sans inscrire de nom de participant dans leur nom.

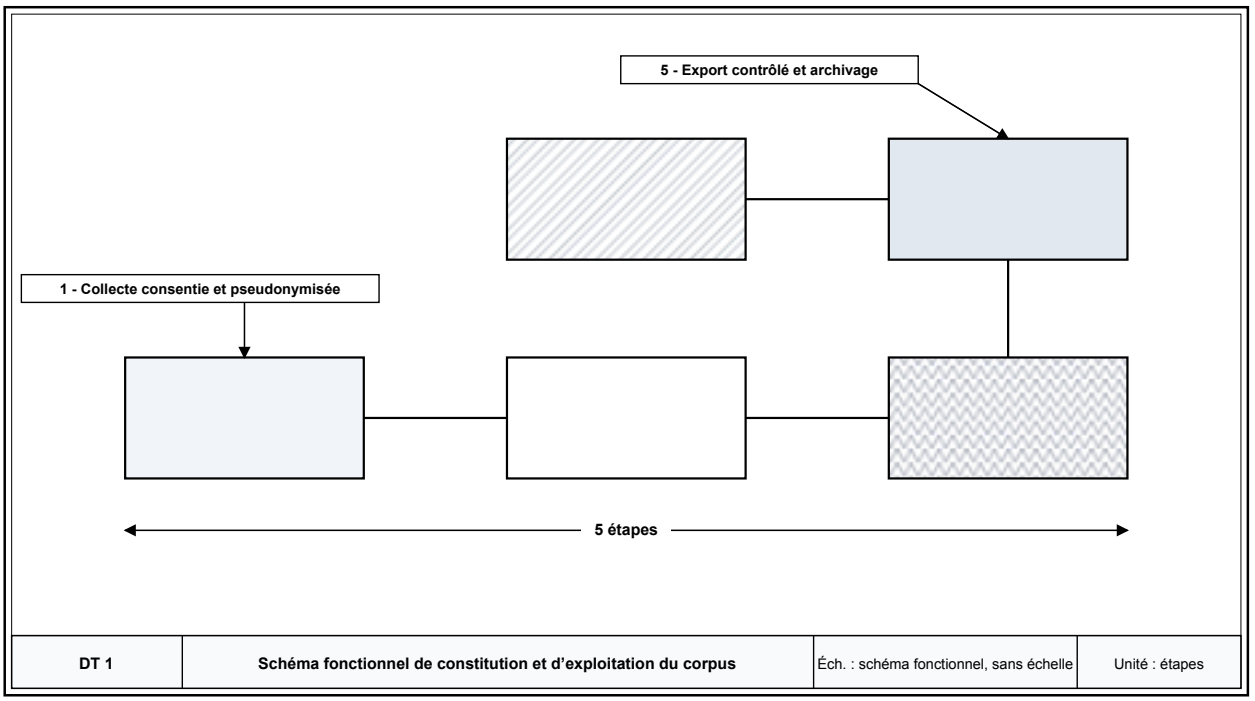
DOSSIER TECHNIQUE ET RESSOURCES

Documents ressources nécessaires à la résolution du sujet

DT 1 : DONNÉES DU CORPUS, CHAÎNE DE TRAITEMENT ET EXTRAITS

Le corpus comprend deux sous-corpus : A, 24 échanges au guichet, 12 minutes d'audio et 7 200 tokens; B, 36 échanges téléphoniques, 18 minutes d'audio et 10 800 tokens. Un token correspond ici à une unité lexicale séparée selon la convention de transcription fournie dans le DT 3. Les comptes vérifiés sont : A, 42 emplois discursifs de *du coup* et 18 de *donc*; B, 36 emplois discursifs de *du coup* et 54 de *donc*. Les occurrences ont été vérifiées manuellement dans leur contexte. Pour comparer des sous-corpus de tailles différentes, une fréquence normalisée peut être calculée par $F = \frac{n}{N} \times 1000$.

Les exemples ci-dessous sont des extraits de travail anonymisés. Les lettres P1 et P2 désignent des locuteurs pseudonymisés. Les crochets signalent une information retirée ou un événement de parole.



Les flèches représentent l'ordre de traitement. Les boîtes décrivent les étapes, non des équipements physiques.

Extraits anonymisés pour l'analyse contextuelle

ID	Extrait	Contexte fourni
A01	P1 : « Il faut prendre un ticket, du coup vous serez appelé. »	Consigne donnée à l'usager
A02	P2 : « Donc, je reviens demain matin ? »	Demande de confirmation
A03	P1 : « Le dossier est complet, du coup on peut l'envoyer. »	Conclusion après vérification
A04	P2 : « J'ai eu un coup au téléphone, du coup j'ai rappelé. »	Emploi de coup dans un groupe nominal
B01	P1 : « Le service ferme à midi, donc appelez avant. »	Information et conseil
B02	P2 : « Du coup, c'est bien le bâtiment B ? »	Reprise et demande de confirmation
B03	P1 : « Il manque une pièce, donc le dossier reste en attente. »	Explication d'une décision
B04	P2 : « J'ai eu du coup une douleur à l'oreille. »	Emploi non nécessairement discursif; contexte à vérifier

Les catégories ne sont pas préattribuées. La forme seule ne suffit pas à décider de la fonction.

DT 2 : CADRE DE COLLECTE, PROTECTION DES DONNÉES ET RÈGLES DE RÉUTILISATION

Pour ce cas d'étude, les participants ont reçu une information sur les objectifs de la collecte, les usages prévus et les modalités de retrait. Le dossier de collecte conserve la version de l'information remise, les autorisations applicables, les dates de collecte et les éventuelles restrictions de réutilisation. Les fichiers audio et les transcriptions peuvent contenir des données permettant d'identifier directement ou indirectement une personne.

Le traitement doit être organisé conformément aux obligations applicables en matière de protection des données personnelles, notamment au règlement (UE) 2016/679 et à la loi Informatique et Libertés modifiée. La pseudonymisation réduit l'identification directe mais ne rend pas, à elle seule, les données anonymes. Toute réutilisation doit être compatible avec l'information fournie, les autorisations et les règles de l'établissement. En cas de doute sur la base de traitement, les destinataires ou la durée de conservation, solliciter le responsable compétent de l'établissement avant toute diffusion.

Mesures opérationnelles prévues pour le cas : accès aux fichiers limité aux personnes autorisées; stockage sur un espace institutionnel protégé; séparation de la table de correspondance des fichiers d'analyse; retrait ou remplacement des noms propres et détails identifiants dans les extraits diffusés; conservation des fichiers selon la durée fixée par le dossier de collecte; signalement des incidents au responsable désigné.

Contrôles documentaires avant réutilisation

Élément à vérifier	Question de contrôle	Risque en cas d'absence de vérification
Information et autorisations	Les usages envisagés sont-ils compatibles avec l'information remise et les autorisations recueillies ?	Réutilisation ou diffusion non prévue
Données identifiantes	Des noms, lieux ou détails contextuels peuvent-ils permettre une identification ?	Atteinte à la confidentialité
Accès et conservation	Qui peut accéder aux fichiers et quelle durée de conservation est prévue ?	Accès excessif ou conservation injustifiée

Les éléments sont à vérifier dans le dossier de collecte et auprès du responsable compétent.

DT 3 : CONVENTIONS D'ANNOTATION, CONTRÔLE DE QUALITÉ ET ORGANISATION DU TRAVAIL

Convention de travail : une occurrence de *du coup* ou de *donc* est retenue dans le décompte des marqueurs discursifs seulement lorsque son emploi contribue à l'organisation du discours, à l'enchaînement des propositions ou à la gestion de l'interaction. Un emploi lexical ou un emploi dont la fonction reste indécidable est marqué « à vérifier » et n'est pas intégré aux comptes validés avant arbitrage. Les décisions sont consignées dans un guide d'annotation comportant la définition des catégories, des exemples positifs et des cas limites.

Contrôle de qualité : un échantillon de 200 unités a été annoté par deux personnes; 172 décisions concordent et 28 divergent. Le nombre d'accords ne fournit pas les distributions par catégorie ni le nombre d'accords attendus par hasard; un coefficient de type kappa ne peut donc pas être calculé à partir des seules données présentes.

Estimation de production pour le cas : 12 minutes d'audio pour A et 18 minutes pour B; temps moyen de transcription, 4,5 minutes de travail par minute d'audio; tarif de calcul, 24 euros par heure de travail. Le calcul demandé porte uniquement sur la transcription, hors charges, équipement, relecture et temps de formation. Pour la prévention au poste informatique, prévoir un aménagement réglable, une posture confortable, un éclairage adapté et des pauses régulières, en tenant compte de l'organisation du travail.

Convention de nommage proposée : code de projet, sous-corpus, numéro séquentiel et version, sans nom de personne, par exemple CORP_A_001_v01. Conserver séparément les métadonnées nécessaires, avec des accès restreints et des versions contrôlées.

Décisions d’annotation et éléments de contrôle

Élément	Valeur	Usage ou limite
Durée audio A	12 min	Base du calcul de temps
Durée audio B	18 min	Base du calcul de temps
Temps de transcription	4,5 min de travail par minute audio	Estimation fournie
Tarif de calcul	24 euros par heure	Hors charges et relecture
Annotation double	172 accords sur 200 unités	Accord observé calculable; kappa non calculable avec les seules données fournies

Les effectifs ci-dessous sont les données disponibles pour cette épreuve.

NE RIEN ÉCRIRE DANS CETTE PARTIE

DR 1 : Document réponse à rendre avec la copie : Questions 2.1 et 2.2 **À RENDRE AVEC LA COPIE**

Complétez directement les cases « ... ». Indiquez les calculs, les unités et les résultats arrondis au centième. Une réponse sans unité ne permet pas de vérifier l'interprétation de la fréquence.

Tableau de calcul des fréquences

Marqueur	Sous-corpus A : occurrences	Sous-corpus A : fréquence / 1 000 tokens	Sous-corpus B : occurrences	Sous-corpus B : fréquence / 1 000 tokens	Total occurrences	Fréquence globale / 1 000 tokens
du coup	42	.....	36	.....	.....	.....
donc	18	.....	54	.....	.....	.....
Total tokens	7 200	.....	10 800	.....	18 000	.....

Fréquences exprimées en occurrences pour 1 000 tokens.

Cadre réservé aux justifications, calculs complémentaires ou schémas du candidat :

.....

.....

.....

.....

.....

.....

.....

NE RIEN ÉCRIRE DANS CETTE PARTIE

DR 2 : Document réponse à rendre avec la copie : Questions 4.1 et 4.3 **À RENDRE AVEC LA COPIE**

Complétez la colonne « Mesure de prévention » avec une action précise et adaptée au risque décrit. Pour la question 4.3, proposez un exemple de nom de fichier conforme aux règles de protection et de versionnage.

Tableau de prévention et de traçabilité

Situation	Risque identifié	Mesure de prévention proposée
Copie d'un fichier audio sur un espace partagé non contrôlé	Accès non autorisé ou réidentification	.....
Séance longue de transcription sans interruption	Fatigue visuelle et inconfort postural	.....
Diffusion d'un extrait contenant un nom propre	Divulgation d'une donnée identifiante	.....
Exemple de nom de fichier anonymisé et versionné	Risque de confusion ou d'inclusion d'une identité dans le nom	.....

Ne faites figurer aucun nom de participant dans les exemples.

Cadre réservé aux justifications, calculs complémentaires ou schémas du candidat :

.....

.....

.....

.....

.....

.....

.....

Copyright © 2027 FormaV®. Tous droits réservés.

FormaV® est une marque déposée auprès de l'INPI sous le numéro 5227897.

Ce document a été élaboré par FormaV® avec le plus grand soin afin d'accompagner chaque apprenant vers la réussite de ses examens. Son contenu (textes, graphiques, méthodologies, tableaux, exercices, concepts, mises en forme) constitue une œuvre protégée par le droit d'auteur.

Toute copie, partage, reproduction, diffusion ou mise à disposition, même partielle, gratuite ou payante, est strictement interdite sans accord préalable et écrit de FormaV®, conformément aux articles L.111-1 et suivants du Code de la propriété intellectuelle. Dans une logique anti-plagiat, FormaV® se réserve le droit de vérifier toute utilisation illicite, y compris sur les plateformes en ligne ou sites tiers.

En utilisant ce document, vous vous engagez à respecter ces règles et à préserver l'intégrité du travail fourni. La consultation de ce document est strictement personnelle.

Merci de respecter le travail accompli afin de permettre la création continue de ressources pédagogiques fiables et accessibles.