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 SL

Sciences du langage

Session 2026 • Code sujet : **26-LIC-SL-EP1** • Durée : **3 heures** • Coefficient : **3,0**
(coefficient indicatif)

Sommaire

Cliquez sur une partie pour y accéder directement

1. Définition de l’objet et cadre de collecte (3 questions) **Page 3**

2. Exploitation quantitative et interprétation (3 questions) **Page 5**

3. Préparation d’un protocole de traitement reproductible (3 questions) **Page 7**

4. Prévention, sécurité et responsabilité dans le travail sur corpus (3 questions) **Page 9**

5. Ce qu’il faut retenir & Questions fréquentes (FAQ) (Synthèse & Q&A Examen) **Page 11**

Sujet officiel (PDF) : https://licencesl.fr/assets/annales/licence-sl/C9 - Mener un travail sur corpus de données langagières/2026/Licence Sciences du Langage - C9 - Mener un travail sur corpus de données langagières - Session 2026 - 1-Sujet.pdf

1. Définition de l’objet et cadre de collecte

La Cellule Corpus compare deux sous-corpus d’échanges de demande d’information, l’un au guichet et l’autre par téléphone, pour étudier les marqueurs du coup et donc. Les candidats formulent une problématique, choisissent une mesure de fréquence adaptée et identifient les précautions liées à la diffusion de données langagières.

Question 1.1. 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.

Problématique : l’emploi discursif de *du coup* et de *donc* varie-t-il entre les échanges au guichet et les échanges téléphoniques ? **Hypothèse vérifiable :** *donc* est plus fréquent pour 1 000 tokens dans le sous-corpus téléphonique B que dans A. Il faut discuter au moins deux variables : la situation et le canal d’interaction, qui diffèrent ensemble, et la composition des échanges, notamment leur durée ou le nombre de locuteurs. La taille des corpus doit aussi être prise en compte par normalisation.

> **✓ RÉPONSE À RETENIR**
>
> Tester si *donc* est plus fréquent dans B, en comparant les fréquences normalisées et en discutant le canal et la composition des échanges.

Question 1.2. 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.

Les comptes bruts ne sont pas directement comparables : A contient 7 200 tokens et B en contient 10 800. Un corpus plus grand peut donc présenter davantage d’occurrences simplement parce qu’il fournit davantage de matière linguistique. La mesure adaptée est une fréquence normalisée, calculée en divisant le nombre d’occurrences par le nombre de tokens, puis en multipliant par 1 000. Son unité est l’occurrence pour 1 000 tokens.

> **✓ RÉPONSE À RETENIR**
>
> Comparer les occurrences pour 1 000 tokens, et non les comptes bruts.

Question 1.3. 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.

Avant de diffuser un extrait ou un fichier, retirer ou remplacer les noms propres et les détails contextuels susceptibles d’identifier une personne : leur présence crée un risque d’atteinte à la confidentialité. Vérifier aussi que les destinataires et les usages prévus sont compatibles avec l’information remise, les autorisations recueillies et les restrictions de réutilisation : sinon, la diffusion peut dépasser ce qui était prévu ou autorisé. La pseudonymisation seule ne garantit pas l’anonymat.

✓ RÉPONSE À RETENIR

Retirer les éléments identifiants et vérifier la compatibilité de la diffusion avec les autorisations et les usages prévus.

2. Exploitation quantitative et interprétation

Les candidats exploitent les occurrences validées de du coup et de donc dans deux sous-corpus de tailles différentes. Ils calculent des fréquences normalisées et globales, complètent le document réponse puis interprètent prudemment les écarts observés.

Question 2.1. 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, en occurrences pour 1 000 tokens, et arrondissez au centième.

La fréquence normalisée est le nombre d'occurrences divisé par le nombre de tokens, puis multiplié par 1 000. Pour *du coup* : A, $42/7200 \times 1000 = 5,83$; B, $36/10800 \times 1000 = 3,33$. Pour *donc* : A, $18/7200 \times 1000 = 2,50$; B, $54/10800 \times 1000 = 5,00$. Les résultats sont en occurrences pour 1 000 tokens, arrondis au centième.

Fréquences normalisées du DR 1

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	5,83	36	3,33		
donc	18	2,50	54	5,00		
Total tokens	7 200		10 800		18 000	

✓ RÉPONSE À RETENIR

Du coup : A 5,83, B 3,33. *Donc* : A 2,50, B 5,00 occurrences pour 1 000 tokens.

Question 2.2. Pour chaque marqueur, calculez le nombre total d'occurrences et la fréquence globale dans les 18 000 tokens réunis. Complétez le DR 1 et détaillez au moins un calcul de fréquence globale.

Marqueur	Total occurrences	Fréquence globale pour 1 000 tokens
Du coup	78	4,33
Donc	72	4,00

Les deux sous-corpus réunissent 18 000 tokens. Pour *du coup* : $(42 + 36)/18000 \times 1000 = 4,33$ occurrences pour 1 000 tokens. Pour *donc* : $(18 + 54)/18000 \times 1000 = 4,00$.

Totaux et fréquences globales du DR 1

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	.....	78	4,33
donc	18	.....	54	.....	72	4,00
Total tokens	7 200	.....	10 800	.....	18 000	.....

✓ RÉPONSE À RETENIR

Du coup : 78 occurrences, soit 4,33 pour 1 000 tokens. *Donc* : 72 occurrences, soit 4,00 pour 1 000 tokens.

Question 2.3. 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.

Du coup est plus fréquent dans A que dans B : 5,83 contre 3,33 occurrences pour 1 000 tokens. À l'inverse, *donc* est plus fréquent dans B : 5,00 contre 2,50. Ces différences décrivent les sous-corpus et peuvent suggérer un lien avec la situation d'interaction, sans démontrer de causalité.

- Les situations de communication et le nombre d'échanges diffèrent.
- Les seules fréquences ne permettent pas d'isoler l'effet du contexte de celui des locuteurs ni de généraliser à d'autres échanges.

✓ RÉPONSE À RETENIR

Du coup : plus fréquent dans A (5,83 contre 3,33). *Donc* : plus fréquent dans B (5,00 contre 2,50), pour 1 000 tokens. Écarts descriptifs, non causaux.

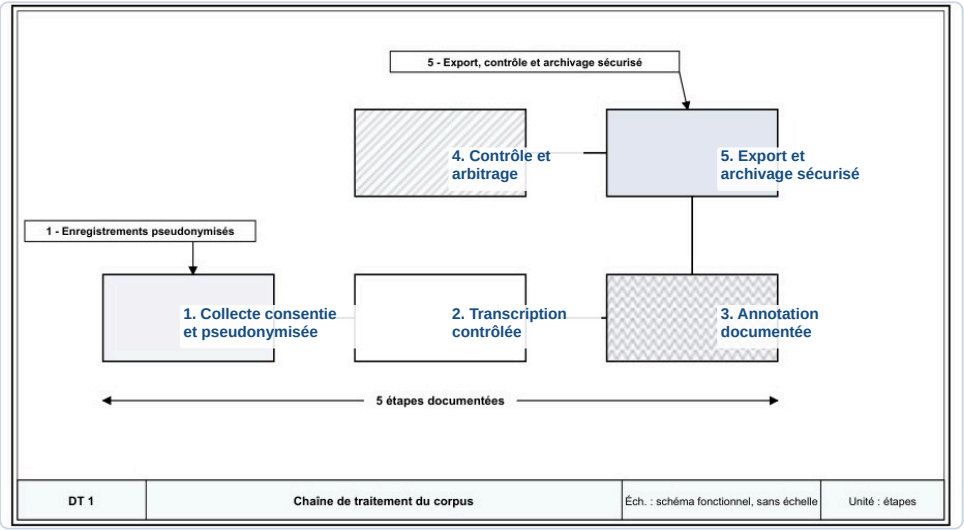
3. Préparation d’un protocole de traitement reproductible

La Cellule Corpus prépare les enregistrements pour l’analyse et veut documenter les étapes de traitement ainsi que les décisions d’annotation. Les candidats décrivent la chaîne de traitement, estiment le temps et le coût de transcription, puis calculent l’accord entre annotateurs et en précisent les limites.

Question 3.1. 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.

1. À réception, vérifier les autorisations, la pseudonymisation et l’intégrité des enregistrements; consigner les fichiers reçus et les anomalies.
2. Stocker les fichiers sur l’espace institutionnel protégé, limiter les accès et séparer la table de correspondance.
3. Transcrire selon les conventions documentées; vérifier un échantillon à l’écoute et conserver la version des conventions.
4. Annoter les emplois en contexte, signaler les cas incertains « à vérifier » et consigner les décisions dans le guide.
5. Contrôler les annotations, arbitrer les divergences, puis exporter et archiver les versions avec un journal des modifications.

Étapes du traitement du corpus



✓ RÉPONSE À RETENIR

Chaîne : vérification, stockage sécurisé, transcription, annotation, contrôle et arbitrage, puis export et archivage versionné.

Question 3.2. **Avec 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 × r et C = Th × ch avec les unités adaptées.**

La durée audio totale vaut 12 + 18 = 30 min. Avec $T = D \times r$, le temps de travail est 30 min audio × 4,5 min de travail/min audio = 135 min de transcription, soit 135 ÷ 60 = 2,25 h. Avec $C = T_h \times c_h$, le coût est 2,25 h × 24 €/h = 54 €. Il s'agit du coût de transcription seul, hors charges, relecture, équipement et formation.

> ✓ **RÉPONSE À RETENIR**
>
> Temps de transcription : 135 min, soit 2,25 h. Coût : 54 €.

Question 3.3. **À partir des 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.**

Le taux d'accord observé se calcule en rapportant les décisions concordantes aux unités annotées : 172 ÷ 200 × 100 = 86 %. Ce pourcentage décrit l'accord brut sur l'échantillon. Il ne mesure pas à lui seul la fiabilité complète : les répartitions par catégorie et le nombre d'accords attendus par hasard ne sont pas connus. Le coefficient de type kappa ne peut donc pas être calculé à partir des données fournies.

> ✓ **RÉPONSE À RETENIR**
>
> Taux d'accord observé : 86 %. La fiabilité complète ne peut être établie avec le seul accord brut.

4. Prévention, sécurité et responsabilité dans le travail sur corpus

Le travail sur le corpus implique la manipulation de fichiers audio et de transcriptions susceptibles de contenir des données identifiantes, ainsi que des séances sur poste informatique. Les candidats proposent des mesures de prévention, vérifient les conditions de réutilisation et définissent un nommage sécurisé des fichiers.

Question 4.1. Pour les trois situations du tableau de la partie 4, proposez une mesure de prévention concrète et proportionnée et complétez le DR 2. Distinguez les mesures relatives à la sécurité des fichiers, à la protection des personnes et à la prévention des risques liés au poste de travail.

Pour chaque situation, choisissez une mesure qui réduit directement le risque indiqué :

Situation	Mesure de prévention
Copie d'un fichier audio	Utiliser l'espace institutionnel protégé et limiter l'accès aux personnes autorisées.
Transcription prolongée	Prévoir des pauses régulières, un poste réglable et un éclairage adapté.
Extrait contenant un nom propre	Retirer ou remplacer le nom et les autres détails identifiants avant diffusion.

Le stockage protégé prévient l'accès indu et la réidentification. Les pauses et l'aménagement du poste réduisent la fatigue visuelle et l'inconfort postural. Le retrait des éléments identifiants protège la confidentialité des personnes. Ces mesures sont adaptées aux trois risques précisés dans le tableau et au cadre opérationnel du dossier technique.

Mesures de prévention complétées

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

✓ RÉPONSE À RETENIR

Stockage institutionnel protégé et accès restreint ; pauses et poste aménagé ; retrait des identifiants avant diffusion.

Question 4.2. 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.

Avant toute réutilisation ou diffusion, vérifiez deux points précis dans le dossier de collecte. D'abord, confrontez les usages envisagés à la version de l'information remise, aux autorisations recueillies et aux éventuelles restrictions : cela évite une utilisation ou une diffusion non prévue. Ensuite, recherchez les noms, lieux et détails

contextuels susceptibles d'identifier une personne, même indirectement. Leur retrait ou leur remplacement dans les extraits diffusés prévient une atteinte à la confidentialité. La pseudonymisation seule ne garantit pas l'anonymat.

✓ RÉPONSE À RETENIR

Vérifier la compatibilité des usages avec l'information et les autorisations, puis repérer les éléments permettant une identification directe ou indirecte.

Question 4.3. Indiquez une règle de nommage ou de versionnage permettant de retrouver les fichiers de travail sans inscrire le nom d'un participant dans leur nom. Proposez un exemple dans le DR 2.

Adoptez une convention stable associant le code du projet, le sous-corpus, un numéro séquentiel et une version. N'inscrivez aucun nom de participant dans le nom du fichier : cette règle facilite le classement et le suivi des modifications sans exposer une identité. Conservez séparément les métadonnées nécessaires, avec des accès restreints et des versions contrôlées. Exemple conforme à la convention du DT 3 : `CORP_A_001_v01`.

Exemple de nom de fichier anonymisé et versionné

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	..... CORP_A_001_v01

✓ RÉPONSE À RETENIR

Convention : projet_sous-corpus_numéro_version, sans nom de participant. Exemple : `CORP_A_001_v01`.

Ce qu'il faut retenir de ce sujet (Session 2026)

Bilan de l'épreuve : Épreuve C9 de sciences du langage, session 2026, code 26-LIC-SL-EP1, coefficient indicatif 3,0. La couverture indique 3 heures, tandis qu'une note introductive mentionne 2 heures : cette incohérence figure dans le sujet. L'étude porte sur du coup et donc dans deux sous-corpus et évalue calcul des fréquences, interprétation, protocole reproductible, annotation et protection des données.

Notions-clés, formules & règles à maîtriser

- Fréquence normalisée : F = (n / N) × 1 000, exprimée en occurrences pour 1 000 tokens.
- Un emploi est retenu comme marqueur discursif selon sa fonction en contexte, et non sur la seule forme du coup ou de donc.
- Les emplois lexicaux et les emplois indécidables ne sont pas intégrés aux comptes validés avant arbitrage.
- Le taux d'accord observé se calcule par accords / unités annotées; un kappa exige aussi les accords attendus par hasard.

Points de vigilance & attentes du jury

- Normalisez les occurrences par le nombre de tokens et indiquez l'unité ainsi que l'arrondi demandé.
- Séparez les constats numériques des interprétations et signalez une limite à la comparaison.
- Pour chaque étape du protocole, indiquez un contrôle de qualité ou une trace utile à la reproductibilité.
- Limitez l'accès aux fichiers, retirez les éléments identifiants des extraits diffusés et excluez tout nom de participant du nommage.

Questions fréquentes sur cette épreuve (Q&A Examen)

Q : Comment calculer une fréquence normalisée dans ce sujet ?

Divisez le nombre d'occurrences validées par le nombre de tokens, puis multipliez par 1 000. Exprimez le résultat en occurrences pour 1 000 tokens et arrondissez au centième, comme demandé.

Q : Pourquoi ne suffit-il pas de compter les formes du coup et donc ?

La forme seule ne distingue pas un marqueur discursif d'un emploi lexical. Examinez le contexte et ne retenez que les emplois qui contribuent à l'organisation du discours, à l'enchaînement des propositions ou à la gestion de l'interaction.

Q : Quelle durée faut-il retenir et comment répartir son temps ?

La couverture indique 3 heures, mais la note introductive mentionne 2 heures. Le sujet comporte donc une incohérence de durée. Les parties valent 5, 6, 5 et 4 points, soit 20 points; prévoyez aussi du temps pour vérifier calculs et unités.

Q : Quelles exigences de réponse faut-il particulièrement respecter ?

Justifiez les calculs, indiquez les unités et distinguez les observations descriptives des interprétations. Documentez les décisions d'annotation et tenez compte des autorisations, des risques d'identification et des règles de sécurité des données.

Sujet officiel associé (PDF) : https://licencesl.fr/assets/annales/licence-sl/C9 - Mener un travail sur corpus de données langagières/2026/Licence Sciences du Langage - C9 - Mener un travail sur corpus de données langagières - Session 2026 - 1-Sujet.pdf
Ressources Licence Sciences du Langage : Toutes les annales & corrigés • Fiches de révision complètes • Simulateur de notes d'examen

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.