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 consulter tous nos sujets d'annales et corrigés de la Licence LLCER, veuillez visiter :
licencellcer.fr/annales-sujets-et-corriges

LICENCE LANGUES, LITTÉRATURES ET CIVILISATIONS ÉTRANGÈRES ET RÉGIONALES • CORRIGÉ DÉTAILLÉ

UE C4 : EXPLOITATION DE DONNÉES

Session 2026 • Code sujet : **26-LIC-LLCE-EP1** • Durée : **2 heures** • Coefficient : **3, indicatif et à confirmer par les modalités de contrôle des connaissances de l'établissement**

Sommaire

Cliquez sur une partie pour y accéder directement

1. **Définition de l'objet d'étude et examen des données** (3 questions) — Page 3
2. **Traitement quantitatif des données** (3 questions) — Page 5
3. **Interprétation et présentation des résultats** (2 questions) — Page 7
4. **Fiabilité, traçabilité et responsabilité de l'analyse** (2 questions) — Page 9
5. **Ce qu'il faut retenir & Questions fréquentes (FAQ)** (Synthèse & Q&A Examen) — Page 10

Sujet officiel (PDF) : https://licencellcer.fr/assets/annales/licence-llcer/C4 - Exploitation de données à des fins d'analyse/2026/Licence Langues, Littératures et Civilisations Étrangères et Régionales - C4 - Exploitation de données à des fins d'analyse - Session 2026 - 1-Sujet.pdf

1. Définition de l'objet d'étude et examen des données

Le corpus comprend quatre sous-corpus écrits répartis selon la langue et le genre, avec leur taille en mots et le nombre d'occurrences d'une marque de première personne. Les documents techniques précisent les formes comptées et les limites du protocole.

Variables et unité d'observation

Question 1.1. À partir des documents DT 1 et DT 2, identifiez la variable de regroupement et la variable mesurée, précisez leur nature et indiquez l'unité d'observation utilisée dans le tableau.

Les variables de regroupement sont la **langue** (anglais ou français) et le **genre** (littéraire ou presse) : ce sont deux variables qualitatives nominales. La variable mesurée est le nombre d'occurrences de la marque de première personne codée dans chaque sous-corpus : c'est une variable quantitative discrète, puisqu'elle dénombre des occurrences entières. L'unité d'observation du tableau est le sous-corpus, repéré par une combinaison de langue et de genre (A, B, C ou D). La taille en mots sert au calcul du taux, mais n'est pas la variable mesurée.

> ✓ RÉPONSE À RETENIR
>
> Regroupement : langue et genre. Mesure : nombre d'occurrences, variable quantitative discrète. Unité d'observation : un sous-corpus.

Comparabilité et normalisation

Question 1.2. Expliquez pourquoi les nombres bruts d'occurrences ne suffisent pas pour comparer les sous-corpus. Donnez la formule de normalisation à appliquer et définissez chacun de ses termes.

Les sous-corpus ont des tailles différentes : un texte plus long peut contenir davantage d'occurrences simplement parce qu'il comporte plus de mots. Il faut donc rapporter le comptage à la taille du sous-corpus, puis exprimer le résultat pour 1 000 mots :

$$t = \frac{n}{N} \times 1000$$

n est le nombre d'occurrences codées, **N** le nombre de mots du sous-corpus et **t** le taux normalisé, exprimé en occurrences pour 1 000 mots. Cette normalisation permet de comparer plus justement les fréquences descriptives entre sous-corpus de tailles différentes.

> ✓ RÉPONSE À RETENIR
>
> Comparer les taux normalisés : $t = (n/N) \times 1000$, en occurrences pour 1 000 mots.

Limite du protocole

Question 1.3. Relevez une limite du protocole de constitution ou de codage des données susceptible d'affecter la comparaison et justifiez votre réponse à partir des documents fournis.

Le protocole ne distingue pas les fonctions pragmatiques des pronoms comptés : il dénombre chaque « I » sujet en anglais et chaque « je » sujet en français, y compris « j' » en français. Ce choix rend le comptage explicite, mais ne permet pas de déduire directement la subjectivité, l'engagement ou la position de l'auteur. De plus, les formes et conventions linguistiques ne sont pas nécessairement équivalentes. Cette limite impose de rester prudent dans l'interprétation des taux, surtout pour une comparaison directe entre langues.

> ✓ **RÉPONSE À RETENIR**
>
> Le comptage ne distingue pas les fonctions pragmatiques : les taux mesurent des occurrences, pas directement la subjectivité ou l'engagement.

2. Traitement quantitatif des données

Les effectifs et les tailles des quatre sous-corpus sont fournis, et les taux doivent être exprimés en occurrences pour 1 000 mots. Les calculs et les résultats demandés sont à présenter avec leurs unités, notamment dans le document réponse DR 1.

Calcul des taux normalisés

Question 2.1. Calculez le taux pour 1 000 mots des sous-corpus A, B, C et D à partir de leurs effectifs et tailles. Présentez les étapes de calcul, complétez le tableau du DR 1 et arrondissez chaque résultat au dixième.

Le taux normalisé rapporte les occurrences au nombre de mots du sous-corpus. Pour chaque ligne, on divise le nombre d'occurrences par la taille du corpus, puis on multiplie par 1 000. Cette opération rend comparables les sous-corpus de tailles différentes. Les calculs reprennent les effectifs du sujet ; chaque résultat est arrondi au dixième et exprimé en occurrences pour 1 000 mots.

Sous-corpus	Calcul	Taux pour 1 000 mots
A	240 ÷ 12 000 × 1 000	20,0
B	90 ÷ 15 000 × 1 000	6,0
C	180 ÷ 10 000 × 1 000	18,0
D	70 ÷ 14 000 × 1 000	5,0

DR 1 complété : taux des sous-corpus

Sous-corpus	Calcul du taux pour 1 000 mots	Taux pour 1 000 mots
A	240/12000 x 1000 = 20,0	20,0
B	90/15000 x 1000 = 6,0	6,0
C	180/10000 x 1000 = 18,0	18,0
D	70/14000 x 1000 = 5,0	5,0

Écarts entre genres, calculés au sein de chaque langue

Langue	Taux littéraire	Taux presse	Écart littéraire moins presse
Anglais			
Français			

✓ RÉPONSE À RETENIR

A : 20,0 ; B : 6,0 ; C : 18,0 ; D : 5,0 occurrences pour 1 000 mots.

Écart entre genres

Question 2.2. Pour chaque langue, calculez l'écart entre le taux du sous-corpus littéraire et celui du sous-corpus de presse. Indiquez le sens de l'écart et son unité, puis complétez le DR 1.

L'écart se calcule en soustrayant le taux de presse du taux littéraire. La comparaison est faite séparément pour chaque langue, à partir des taux arrondis au dixième de la question précédente. Un écart positif signifie que le taux littéraire observé est supérieur à celui de la presse. L'unité de l'écart est l'occurrence pour 1 000 mots.

Langue	Taux littéraire	Taux presse	Écart littéraire moins presse
Anglais	20,0	6,0	+14,00
Français	18,0	5,0	+13,00

Dans les deux langues, l'écart est positif : le taux du sous-corpus littéraire est plus élevé.

DR 1 complété : écarts entre genres

Écarts entre genres, calculés au sein de chaque langue

Langue	Taux littéraire	Taux presse	Écart littéraire moins presse
Anglais	20,0	6,0	+14,00
Français	18,0	5,0	+13,00

✓ RÉPONSE À RETENIR

Anglais : +14,00 ; français : +13,00 occurrences pour 1 000 mots. Le taux littéraire est supérieur dans les deux cas.

Rapport entre les taux

Question 2.3. Pour chaque langue, calculez le rapport entre le taux littéraire et le taux de presse, arrondissez au centième et interprétez ces rapports sans conclure à une différence statistiquement significative.

Le rapport demandé est le taux littéraire divisé par le taux de presse, au sein de chaque langue. Il exprime combien de fois le taux littéraire est celui de la presse. Les valeurs ci-dessous sont arrondies au centième et décrivent uniquement les sous-corpus observés. Elles ne constituent pas un test statistique et ne permettent pas, à elles seules, de conclure à une causalité ou à une différence statistiquement significative.

1. Anglais : 20,0 ÷ 6,0 = **3,33**. Le taux littéraire est environ 3,33 fois celui de la presse.
2. Français : 18,0 ÷ 5,0 = **3,60**. Le taux littéraire est 3,60 fois celui de la presse.

✓ RÉPONSE À RETENIR

Rapport littéraire/presse : anglais 3,33 ; français 3,60. Ces rapports sont descriptifs et ne démontrent pas une différence statistiquement significative.

3. Interprétation et présentation des résultats

Les résultats quantitatifs doivent être interprétés en tenant compte des limites du corpus et du protocole de comptage. Les comparaisons entre genres se font au sein de chaque langue, et toute comparaison entre langues doit signaler que les formes codées et les corpus ne sont pas nécessairement équivalents.

Conclusion sur les contrastes observés

Question 3.1. Rédigez une conclusion de 80 à 120 mots décrivant le contraste observé entre les deux genres pour chacune des langues. Distinguez explicitement le constat quantitatif de son interprétation linguistique et respectez les limites du corpus.

Constat quantitatif : en anglais, le taux est de 20 occurrences pour 1 000 mots en littérature, contre 6 dans la presse, soit un écart de 14. En français, les taux sont respectivement de 18 et 5, soit un écart de 13 occurrences pour 1 000 mots. La marque codée est donc plus fréquente dans les sous-corpus littéraires des deux langues. **Interprétation linguistique :** ce résultat peut suggérer une présence plus fréquente de la première personne dans ces textes, mais le comptage ne mesure ni la subjectivité ni l'engagement. La taille et la composition des échantillons, ainsi que les formes codées (« I » et « je »), limitent la comparaison, surtout entre langues. L'écart ne prouve ni causalité du genre ni différence statistiquement significative.

> ✓ **RÉPONSE À RETENIR**
>
> Dans les deux langues, le taux observé est supérieur en littérature. Ce constat descriptif ne prouve ni causalité du genre ni différence statistiquement significative.

Représentation graphique

Question 3.2. Proposez une représentation graphique adaptée pour comparer les quatre taux. Précisez le type de graphique, les variables à placer sur chaque axe, l'unité, une légende nécessaire et un titre informatif. Indiquez une précaution de lecture à faire figurer sous le graphique et complétez le DR 2.

Sur le DR 2, recommander un diagramme en barres : l'axe horizontal présente les sous-corpus A, B, C et D ; l'axe vertical indique le taux en occurrences pour 1 000 mots. Le titre précise qu'il s'agit de la fréquence de la marque de première personne par sous-corpus. La légende distingue les genres : A et C, littérature ; B et D, presse. Ajouter une précaution : comparer les genres au sein de chaque langue, car les formes codées et les corpus ne sont pas nécessairement équivalents entre langues, et les taux restent descriptifs.

DR 2 complété : spécification du graphique

Élément	Réponse du candidat
Type de graphique	Diagramme en barres
Variable de l'axe horizontal	Sous-corpus : A, B, C, D
Variable et unité de l'axe vertical	Taux : occurrences / 1 000 mots
Titre informatif	Taux de la marque de 1re personne par sous-corpus
Légende ou précaution de lecture	Littéraire : A, C ; presse : B, D. Comparer les genres par langue ; taux descriptifs.

✓ RÉPONSE À RETENIR

Diagramme en barres : A 20,0 ; B 6,0 ; C 18,0 ; D 5,0 occurrences pour 1 000 mots.

4. Fiabilité, traçabilité et responsabilité de l'analyse

Les documents techniques décrivent les règles de préparation et de comptage ainsi que les limites d'interprétation et les principes de diffusion. Les réponses doivent proposer des mesures concrètes de reproductibilité et des précautions de diffusion adaptées à cette analyse.

Reproductibilité et vérification

Question 4.1. Proposez deux actions permettant à une autre personne de reproduire le comptage et de vérifier les résultats. Pour chacune, précisez l'information ou la procédure à conserver.

1. Conserver le protocole documenté : critères de sélection des textes, étapes de préparation en texte brut, formes comptées (« l », « je » et « j' ») et exclusions des métadonnées, titres et commentaires. Cela permet de refaire le comptage selon les mêmes règles.
2. Archiver les données de comptage et le détail des calculs : nombre de mots, occurrences par sous-corpus, paramètres de comptage, formule appliquée et arrondis. Une autre personne peut ainsi recalculer les taux et vérifier les résultats, dans le respect des droits et de la confidentialité.

> **✓ RÉPONSE À RETENIR**
>
> Conserver le protocole complet, les critères de sélection, les données de comptage et le détail des calculs.

Précautions avant diffusion

Question 4.2. Identifiez deux précautions à prendre avant de diffuser les textes ou les résultats détaillés. Reliez chacune à un risque pertinent, par exemple les droits sur les textes, la présence de données personnelles ou une généralisation abusive à partir d'échantillons limités.

Avant toute diffusion, vérifier les droits applicables aux œuvres et les conditions d'accès aux sources ; ne publier les textes ou extraits que si leur diffusion est autorisée. Examiner aussi les textes et métadonnées pour repérer d'éventuelles données personnelles, puis limiter leur diffusion et appliquer les règles de protection pertinentes.

- Limiter les extraits et retirer ou protéger les informations personnelles pour éviter une divulgation non autorisée.
- Présenter les résultats comme descriptifs, préciser la taille et la composition limitées des échantillons et éviter d'attribuer les écarts au genre sans analyse complémentaire.

> **✓ RÉPONSE À RETENIR**
>
> Vérifier les droits et protéger les données personnelles ; contextualiser les résultats pour éviter les généralisations abusives.

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

Bilan de l'épreuve : Cette proposition d'évaluation d'établissement de l'UE C4 LLCER, session 2026, porte le code 26-LIC-LLCE-EP1 et dure 2 heures ; le coefficient 3 est indicatif et à confirmer localement. Elle étudie les marques de première personne dans quatre sous-corpus anglais et français, littéraires et de presse. L'épreuve évalue la normalisation des fréquences, leur comparaison, l'interprétation critique, la représentation graphique et la reproductibilité.

Notions-clés, formules & règles à maîtriser

- Le taux normalisé est t = (n/N) × 1 000, où n désigne les occurrences codées et N le nombre de mots du sous-corpus.
- Les taux obtenus sont A : 20,0, B : 6,0, C : 18,0 et D : 5,0 occurrences pour 1 000 mots.
- L'écart littéraire moins presse est de 14,0 occurrences pour 1 000 mots en anglais et de 13,0 en français.
- Les rapports littéraire/presse sont 3,33 en anglais et 3,60 en français ; ces mesures descriptives ne prouvent pas une différence statistiquement significative.

Points de vigilance & attentes du jury

- Ne comparez pas les effectifs bruts sans tenir compte des tailles différentes des sous-corpus.
- Indiquez les unités et respectez les arrondis demandés : taux au dixième et rapports au centième.
- Distinguez les résultats observés de leur interprétation et ne présentez pas un écart comme une preuve de causalité.
- Limitez les comparaisons directes entre langues : les formes comptées et la composition des corpus peuvent différer.

Questions fréquentes sur cette épreuve (Q&A Examen)

Q : Comment calculer un taux pour 1 000 mots ?

Divisez le nombre d'occurrences codées par le nombre de mots du sous-corpus, puis multipliez le résultat par 1 000. Le taux s'exprime en occurrences pour 1 000 mots et doit être arrondi au dixième dans cette épreuve.

Q : Que faut-il compter pour les pronoms français et anglais ?

Le protocole compte le pronom sujet I en anglais et je en français. La forme élidée j' est comptée comme une occurrence de je. Les occurrences présentes dans les métadonnées, les titres de fichiers et les commentaires éditoriaux sont exclues.

Q : Comment gérer le temps et les points de l'épreuve ?

La durée indiquée est de 2 heures et les quatre parties totalisent 20 points. Aucun temps par partie n'est prescrit : répartissez-le selon les points, en gardant une période pour vérifier les calculs, les unités et les documents réponses.

Q : Un écart de taux suffit-il à prouver que le genre explique la différence ?

Non. Les taux décrivent uniquement les sous-corpus étudiés et ne démontrent ni causalité ni significativité statistique. La taille, la composition et la période des échantillons, ainsi que les conventions linguistiques, limitent la portée des conclusions.

Sujet officiel associé (PDF) : https://licencellcer.fr/assets/annales/licence-llcer/C4 - Exploitation de données à des fins d'analyse/2026/Licence Langues, Littératures et Civilisations Étrangères et Régionales - C4 - Exploitation de données à des fins d'analyse - Session 2026 - 1-Sujet.pdf

Ressources Licence Langues, Littératures et Civilisations Étrangères et Régionales : 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.