

## Contrôle des connaissances : Linguistique de corpus

Répondez aux questions qui suivent.

### Exercice 1 (sur 6 points)

Donner la définition de *corpus* qui a été proposée par Sinclair. Puis, à partir de cette définition et des notions comme *échantillon* et *représentativité*, expliquer ce qui distingue les *corpus de référence* des *corpus de spécialité*.

### Exercice 2 (sur 8 points)

- Certains auteurs proposent de ne pas recourir à la notion de *phrase* pour segmenter l'oral en unités linguistiques, mais utilisent les unités ou notions suivantes : le noyau, le préfixe, le suffixe et le postfixe.  
Expliquer ce que sont les unités suivantes, en vous appuyant sur des exemples extraits du texte qui suit : le *noyau*, le *préfixe* et le *suffixe*.
- Segmenter le texte qui suit en phrases ou en *noyaux*, *suffixes* et *préfixes*. Ne pas hésiter à commenter la segmentation proposée, surtout en cas de difficultés. Attention, aucune segmentation n'est donnée dans le texte.

**Texte (extrait de ACSYNT, LAUE) :**

L : en dehors de l'école|on était tout le temps dehors|sauf quand c'était euh l'hiver|mais euh|je me souviens que l'été on était tout le temps dehors|que mes parents ils avaient du mal à nous récupérer|souvent pour les repas|on rentrait parce qu'on avait faim quoi|mais [rires] souvent|on était tout le temps par monts et par vaux|on allait construire des cabanes dans la forêt de Saint Saint-Affrique|on était tout le temps dehors|oui même y compris après dîner|le soir|on ressortait euh jusqu'à dix heures onze heures|on avait la permission de de faire les idiots dans le villages.

### Exercice 3 (sur 6 points)

En s'appuyant sur le mémo qui suit, utiliser les expressions régulières pour rechercher dans un document écrit (corpus ou archive) les éléments mentionnés dans les questions.

Mémo :

- Les caractères littéraux sont interprétés comme ils sont écrits : la séquence *chat* permet de retrouver des formes comme chat, chats, chatons, chatte, achat, etc.
- Les débuts et fin de mot sont indiquées par \b. La séquence *\bours\b* permet de rechercher uniquement ours et d'exclure ourson, bourse, etc.
- Les crochets servent à délimiter une classe ou un ensemble de caractères. A l'intérieur des crochets, ils sont mentionnés les uns à la suite des autres, sauf s'il s'agit d'un intervalle comme 'tout nombre de 0 à 9' (noté alors [0-9]). La séquence *él[éeè]* permet de trouver des formes comme élève, élevons, élévation.
- Le point d'interrogation à la suite d'un caractère ou d'une suite de caractères indique 0 à 1 occurrence de ce caractère.

1. Rechercher les déterminants partitifs du, des, de. Dire si la requête proposée risque de faire sortir des éléments qui ne sont pas des déterminants partitifs ? Si oui, lesquels. Pour répondre, regarder le texte de la question précédente.
2. Rechercher les formes verbales au conditionnel. Expliquer clairement sa démarche en donnant les terminaisons recherchées avant d'écrire l'expression régulière.

⌘ Les parenthèses et les tirets verticaux permettent de mentionner des alternatives. La séquence *(ons|ez|ent)* permet de trouver les formes contenant *ons* comme *mangeons* ou *bonsoir*, celles contenant *ez* et celles contenant *ent*.