X3IT060
Corpus

Université de Nantes
2019/2020

Examen n° 1

Durée : 1 H 15
Documents autorisés : une feuille A4 recto-verso - Calculette
Novembre 2019

Exercice 1.1 (Définition de corpus)

Quel est l'intérêt de ne considérer comme éléments constitutif d'un corpus que des textes intégraux ? À l'inverse, quels peuvent être les problèmes ?

Exercice 1.2 (Phrasèmes)

a) Soit l'expression *prendre une veste* qui peut selon les contextes avoir un emploi idiomatique dans le sens "essuyer un échec" ou un emploi littéral. Donner un exemple de contexte où l'emploi de cette expression ne peut pas induire une interprétation idiomatique.
b) Les expressions ci-dessous sont-elles des expressions figées ou des collocations. Justifier votre réponse. Préciser pour les collocations, ses caractéristiques : base et collocatif, type syntaxique, relation syntaxique entre base et collocatif et type sémantique.

- *tirer son épingle du jeu*
- *traduire en justice*
- *tourner la page*
- *faire face*
- *frôler la mort*
- *perdre la raison*

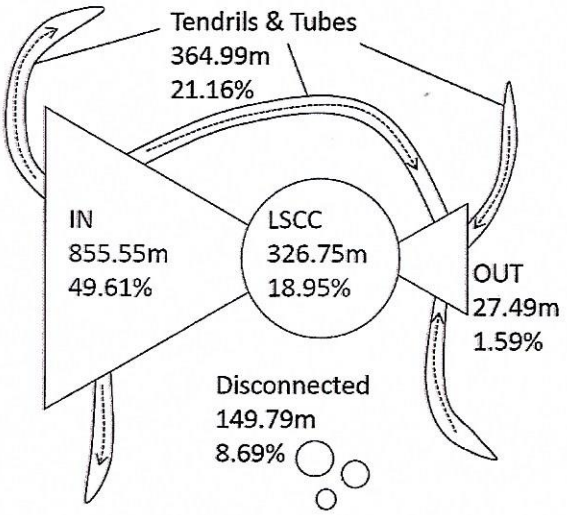
c) Quels sont les phrasèmes ci-dessus ambigus, i.e. acceptant un emploi idiomatique et un emploi littéral ? Expliquer.

Exercice 1.3 (NLTK - Python)

- Présenter en quelques mots les corpus Reuters et TIMIT distribués par NLTK
- Qu'est-ce qui distingue les catégories du corpus Brown et du corpus Reuters ?
- Les catégories du corpus Brown et celles du corpus Reuters peuvent-elles être assimilées aux genres tel que vu en cours ?
- Expliquer la différence d'usage entre ces deux méthodes : `FreqDist()` et `ConditionalFreqDist()`

Exercice 1.4 (Web comme corpus)

(i) Voici un schéma représentant la structure du web en 2014. Par rapport à celle de 2000 en forme de papillon, expliquer ce qu'elle induit sur la construction de corpus à partir du web, en particulier en fonction des méthodes de crawling BootcCat ou moissoneur.



(ii) Bjoneborn et Ingwersen (2004) et autres chercheurs considèrent qu'ils y a environ dix niveaux de granularité pour caractériser le web. La liste ci-dessous les décrit du plus petit au plus important :

1. bit
2. caractère
3. mot
4. bloc
5. page
6. sous-site
7. site (équivalent au client/hote)
8. domaine (équivalent au serveur)
9. domaine général (Top Level Domain)
10. domaine national
11. web global

Rastier (2001) a proposé une typologie comportant sept éléments distinctifs pour caractériser les corpus :

1. langues
2. discours : chaque langue connaît des usages propres à des types de pratiques sociales
3. genres : chaque discours compte un nombre déterminé de genres, dont la typologie rend compte de la diversité externe des textes
4. textes : la typologie des textes traite de la diversité interne des genres
5. sections et configurations : parties de textes délimitées par des critères d'expression ou par des parties de contenu
6. morphologies : la parenté des textes, indépendemment des genres. Chaque texte possède un vocabulaire de "formes sémantiques"
7. usages génériques ou styles : ils différencient des classes d'usagers ou des "styles"

Si l'on considère le web comme un corpus, quels niveaux du web pourraient être mis en correspondance avec les niveaux de caractérisation de Rastier ? Dans le cas ou la correspondance n'existe pas, comment y palier ? Discuter en particulier les genres, discours et langues.

Exercice 1.5 (Annotation)

(i) En quoi la délimitation des unités à annoter est-elle un facteur de complexité de l'annotation. Illustrer sur une annotation en entités nommées sur le texte ci-dessous :

C'est sur les hauteurs d'Antananarivo, dans l'enceinte du palais d'Andafiavaratra, que s'est tenue, lundi 18 novembre, la réunion de la commission mixte franco-malgache qui doit décider de l'avenir des îles Eparses. Ces cinq îlots administrés par la France sont revendiqués par Madagascar depuis les années 1970. L'idée de cette commission avait été proposée par le président Hery Rajaonarimampianina, le prédécesseur d'Andry Rajoelina.
Côté malgache, la délégation était conduite par le premier ministre et ministre des affaires étrangères par intérim, Christian Ntsay. Dans ses rangs, on compte notamment Raymond Ranjeva, ancien vice-président de la Cour internationale de justice et actuel président de l'Académie malgache.

(ii) Lors d'une annotation, deux annotateurs A et B examinent des phrases et doivent indiquer si oui ou non celle-ci contient une date. Soit la matrice de confusion ci-dessous résumant illustrant les résultats de cette annotation double menée par les annotateur A et B.

		A	**A**
		Oui	Non
B	Oui	10	0
B	Non	3	4

1. Quel est le pourcentage d'accord ?
2. Quel est le pourcentage de désaccord ?
3. Quel est l'accord inter-annotateur (accord observé) ?
4. Quel est l'accord attendu (S de Scott, π de Scott, κ de Cohen) en rappelant ce qui distingue ces accords ? Interpréter ces mesures.

(ii) Pour améliorer l'accord, on rajoute une catégorie "indécis".

		A	**A**	**A**
		Oui	Non	Indécis
B	Oui	10	0	0
B	Non	3	4	0
B	Indécis	0	0	0

Recalculer les métriques d'accord comme pour (i). Commentez.