X3IT060
Corpus

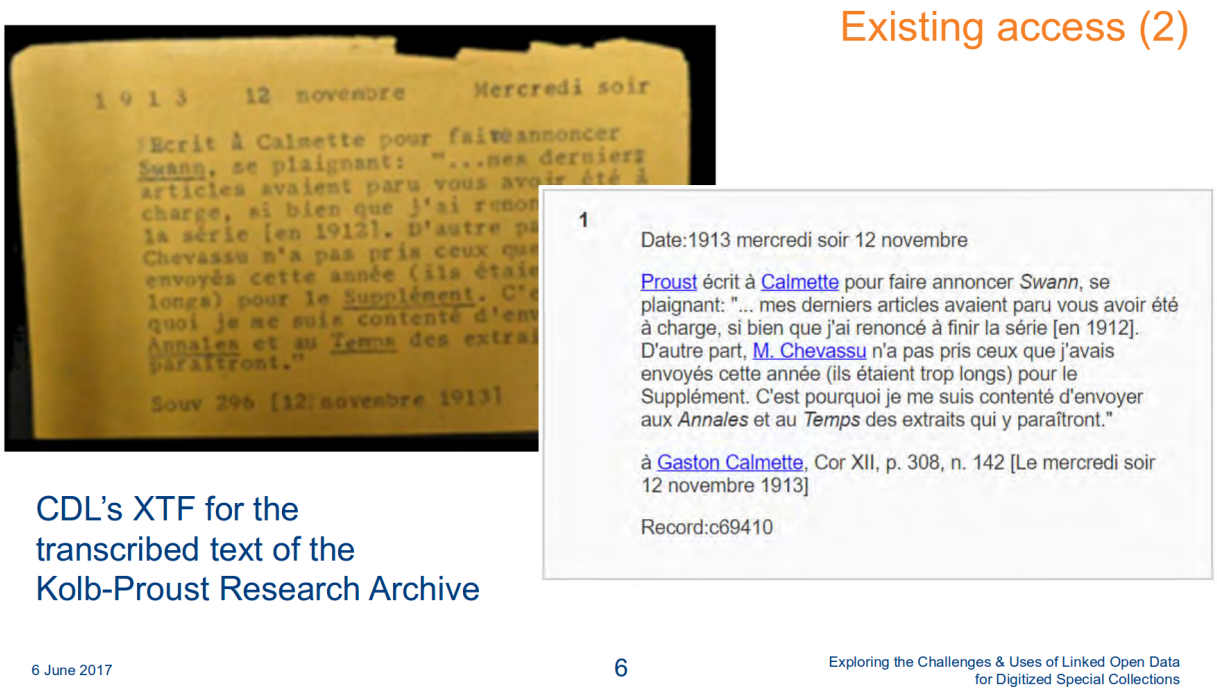
Université de Nantes
2018/2019

Examen n° 1

Durée : 1 H 30
Documents autorisés : une feuille A4 recto-verso - Calculette
Décembre 2018

Exercice 1.1 (Métadonnées)

Soit le texte retranscrit à partir du texte dactylographié extrait de l'archive de textes Kohn-Proust :



---

1. Renseigner le jeu de méta-données DUBLIN CORE pour ce document (Title, Author, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights).
2. Quel sont les éléments qui pourraient être renseignés automatiquement ? Comment ?
3. Rajouter en étant conforme au jeu de méta-données DUBLIN CORE, les éléments suivants :
   — le genre (à proposer),
   — l'intention de communication (à choisir parmi : Information, Discussion, Instruction, Inventaire, Loi, Promotion, Reportage, Autre),
   — le registre de langue (à choisir parmi : familier/courant/soutenu),
   — la caractérisation de l'auteur (à choisir parmi : un auteur/plusieurs coauteurs/collectif),
   — son audience (catégorie du destinateur et du destinataire à choisir parmi : spécialiste/grand public).
   Justifier vos réponses.

Exercice 1.2 (Phrasèmes)

Identifier dans le texte les phrasèmes demandés ci-dessous en justifiant votre réponse. En particulier, vous préciserez la catégorie morphosyntaxique du phrasème ainsi que celle de ses constituants. Pour les collocations, vous préciserez en outre les caractéristiques suivantes : base et collocatif, type syntaxique et type sémantique. Vous diversifierez au mieux vos exemples.

— 1 expression figée
— 2 mots composés
— 3 collocations

Mourir de chaud, un risque pour 30 % de la population mondiale
Sans une réduction drastique des gaz à effet de serre, les trois quarts des habitants de la planète seraient exposés à des vagues de chaleur potentiellement mortelles à la fin du siècle. Alors qu'une partie de la France est placée en vigilance orange pour la canicule, une étude se penche sur le risque de « mourir de chaud », au sens propre. Publiée en ligne lundi 19 juin dans la revue Nature Climate Change, elle conclut que ce danger guette aujourd'hui près d'un individu sur trois dans le monde. Une proportion qui pourrait grimper à trois sur quatre à la fin du siècle, si les émissions de gaz à effet de serre se poursuivent à leur rythme actuel.
Pour poser ce diagnostic, une équipe américano-britannique de dix-huit chercheurs, dont la plupart travaillent à l'université de Hawaï, a compilé la littérature scientifique documentant les cas de mortalité supplémentaire associée à des vagues de chaleur, entre 1980 et 2014. Elle en a identifié 783, observés dans 164 villes de 36 pays. Parmi eux figurent la canicule de l'été 2003, à l'origine de 70 000 morts excédentaires en Europe, dont environ 20 000 en France, et près de 5 000 à Paris, celle de 2010, à laquelle sont imputés 55 000 décès supplémentaires en Russie, dont près de 11 000 à Moscou, ou celle qui avait frappé Chicago en juillet 1995, responsable de plus de 700 morts.
Les auteurs ont ensuite croisé ces données avec les paramètres climatiques enregistrés lors de ces épisodes : température de l'air, taux d'humidité relative, ensoleillement, vitesse du vent... Ils en ont déduit que le facteur déterminant, pouvant altérer la capacité de thermorégulation de l'organisme humain et provoquer un état d'hyperthermie, était le couple température-humidité, cette dernière renforçant la chaleur ressentie. Ils ont alors calculé un seuil à partir duquel l'association de ces conditions ambiantes peut devenir fatale. Les chercheurs n'affirment évidemment pas que le dépassement de ce seuil conduit à un trépas inéluctable, mais simplement qu'il expose à un « coup de chaud » potentiellement mortel. Différentes parades peuvent en effet être mises en œuvre pour éviter une telle extrémité, allant de l'équipement individuel en système de climatisation jusqu'à la politique publique de prévention. A l'aune de ce critère, l'équipe a établi qu'en 2000, le seuil fatidique de température et d'humidité a été franchi, pendant au moins vingt jours, sur environ 13 % de la surface continentale de la planète, abritant 30 % de la population mondiale.

Exercice 1.3 (Définition de corpus)

Qu'apporte la définition d'un corpus de Habert (2000) par rapport à celle de Sinclair (1996) ? Et celle de Rastier (2002) par rapport à celle de Habert (2000) ? Ces définitions sont-elles complémentaires, contradictoires ?

Exercice 1.4 (Genres du web)

Expliquer pourquoi il est difficile de proposer une typologie des genres du web a priori. Pour illustrer vos réponses vous pouvez vous appuyer sur la typologie suivante établie par des journalistes dans leur tache d'écriture d'articles.
article ; page gouvernementale ; page d'accueil ; index ; page d'information liste ; page principale ; moteur de recherche ; page de recherche ; page avec les résultats de recherche ; carte (géographique) ; résumé ; table des matières ; magazine ; page "qui sommes-nous ?" ; page avec des publicités ; blog ; page d'accueil d'une entreprise ; page d'accueil d'un organisme ; page encyclopédique ; FAQ ; lettre ; liste de liens ; page de navigation ; PDF ; questions-réponses ; termes et conditions ; archive de résumés ; sommaire ; page de magazine ; compte-rendu de réunion ; page éducative ; couverture ; journal

Exercice 1.5 (Calul et interprétation d'accords inter-annotateurs)

Soit la matrice de confusion ci-dessous illustrant les résultats d'une annotation double menée par les annotateurs A et B sur deux catégories oui et non.

		A	**A**
		Oui	Non
B	Oui	30	20
B	Non	10	40

1. Quel est l'accord inter-annotateur (accord observé) ?
2. Quels sont les valeurs des mesures inter-annotateurs (S de Scott, π de Scott, κ de Cohen) en rappelant les différentes hypothèses des accords attendus ?
3. Interpréter ces mesures.

Exercice 1.6 (Web comme corpus)

1. Préciser quels sont les éléments-graines dans les deux principales méthodes de moissonnage du web. Donner deux avantages et deux inconvénients de chacune de ces méthodes.
2. Expliquer comment interpréter une faible probabilité ($< 0{,}01$) au test du X^2 lors d'une comparaison de liste de fréquences entre deux corpus.
3. Calculer la corrélation de Spearman ρ entre les corpus ci-dessous. Renseigner les colonnes vides dans le tableau ci-dessous. Interprétez. La feuille remplie sera jointe à votre copie d'examen. N'oubliez pas d'y indiquer votre nom.

	Fréquences		Rangs		Calculs	
mots	Corpus1	Corpus2	Corpus1	Corpus2	D	D^2
de	6781719	2446328				
,	5627749	2564809				
la	3613946	2653365				
.	3574395	2956662				
que	2963992	3579032				
y	2642241	3614049				
en	2562028	5633555				
el	2450353	6802262				
a	1885112	1882813				
los	1597103	1000106				
del	1173860	1054924				
se	1139311	1143202				
las	10554729	1172623				
un	1001556	1603537				