

Corpus : définition et droits

Karën Fort

karen.fort@univ-lorraine.fr / <https://members.loria.fr/KFort/>

Quelques sources d'inspiration

- ▶ Corpus Linguistics [McEnery and Wilson, 1996],
- ▶ Cours de Cédric Faron et Anne Catherine Simon (Université de Louvain) : Méthodologie de l'analyse de corpus en linguistique.
- ▶ Wikipédia sur les droits d'auteur

Pourquoi en parler ?

utilisationS de corpus dans beaucoup de vos cours et en TAL en général

- ▶ définir l'objet que vous manipulez
- ▶ connaître les possibilités et les limites de son utilisation

Qu'est-ce qu'un corpus ?



Définition

Corpus

*A corpus is a collection of **pieces** of language that are **selected** and **ordered** according to **explicit** linguistic [and/or extra-linguistic] **criteria** in order to be **used** as a sample of the language [Sinclair, 1996]*

«Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage.»

Introduction

Corpus et corpus

Corpus et droits

Pour finir

?

texte

?

texte parole

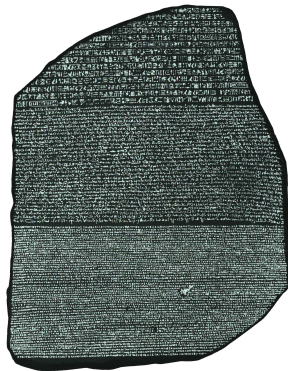
?

texte parole musique

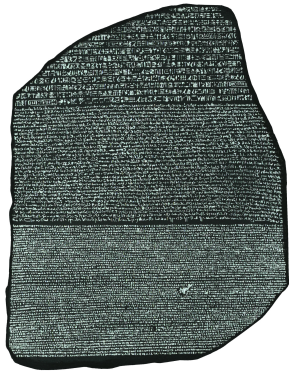
?

texte parole musique vidéo

Monolingue vs Multilingue



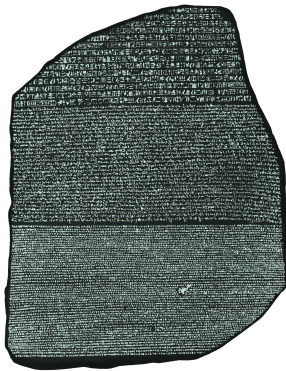
Monolingue vs Multilingue



«La pierre de Rosette est un fragment de stèle gravée de l'Égypte antique portant trois versions d'un même texte [...]. L'inscription qu'elle comporte est un décret promulgué à Memphis par le pharaon Ptolémée V en 196 av. J.-C. Le décret est écrit en deux langues (égyptien ancien et grec ancien) et trois écritures : égyptien en hiéroglyphes, égyptien démotique et alphabet grec.»

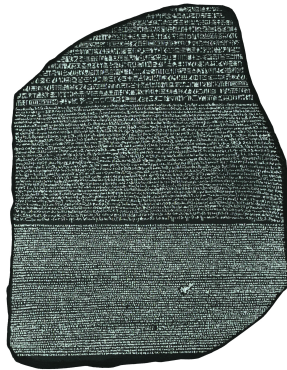
[Wikipédia, 22 sept. 2015]

Monolingue vs Multilingue



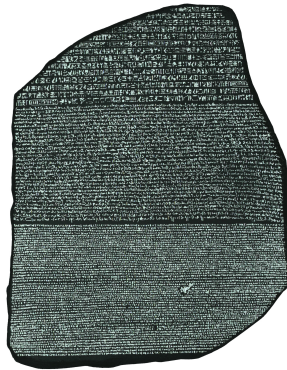
aligné vs comparable

Monolingue vs Multilingue



1 ou 2 (3) corpus ?

Monolingue vs Multilingue



1 ou 2 (3) corpus ? dépend de l'**application** !

Finis vs Ouverts vs Dynamiques [Baude, 2007]

- ▶ **Finis / clos** : construit une fois pour toute comme un corpus « complet » [Corpus de référence du français parlé 1 ; Delic 2004]
- ▶ **Ouverts** : construit pour intégrer de nouvelles données, qu'elles soient prévues ou non [Web, presse en ligne]
- ▶ **Dynamiques** : sous-catégorie des corpus ouverts, qui inclut les corpus de surveillance [COBUILD] et les corpus réservoirs [VALIBEL]

Exhaustif vs Représentatif vs Équilibré vs de référence [Baude, 2007]

- ▶ Exhaustif : corpus fini contenant tous les textes correspondant à un usage particulier (d'un auteur, par exemple)
- ▶ Représentatif : notion vague, par genre, par échantillon sociologique, par situation communicationnelle
- ▶ Équilibré : échantillons de textes (Brown corpus)
- ▶ de référence : construit pour fournir des informations sur une langue, de grande taille et diversifié

Données brutes vs données construites [Baude, 2007]

Données **naturelles** vs données **créées** (interviews, etc)

Petite taille vs grande taille [Baude, 2007]

Qu'est-ce qu'être **gros** pour un corpus ?

Collection organisée de données vs banque de données [Baude, 2007]

Sélection ?



[Joe Crawford from Moorpark, California, USA]

Sacs de mots vs collections de textes [Baude, 2007]

- ▶ textes **structurés** ou liste de mots *indépendants* ?
- ▶ textes **complets** ou parties de textes (échantillons) ?

Brut vs annoté [Baude, 2007]

Est-ce que la **transcription** (de la parole) est une annotation ?

Court terme vs long terme [Baude, 2007]

- ▶ corpus créé pour **un** projet de recherche
- ▶ corpus utilisable dans **plusieurs** projets de recherche
- ▶ corpus proposant des annotations **partageables** (standards)

Conclusion

- ▶ variété de points de vue
- ▶ pas que du texte !

Introduction

Corpus et corpus

Corpus et droits

Nul n'est censé ignorer la loi

Licences libres

Pour finir

Introduction

Corpus et corpus

Corpus et droits

Nul n'est censé ignorer la loi

Licences libres

Pour finir

Trouver des textes

Faites la recherche suivante dans votre navigateur préféré :

"hitchhicker's guide to the galaxy pdf"

Que trouvez-vous ? Avez-vous le droit de l'utiliser ?

Trouver des textes

Faites la recherche suivante dans votre navigateur préféré :

"victor hugo les misérables pdf"

Que trouvez-vous ? Avez-vous le droit de l'utiliser ?

Droit d'auteur

dans l'UE

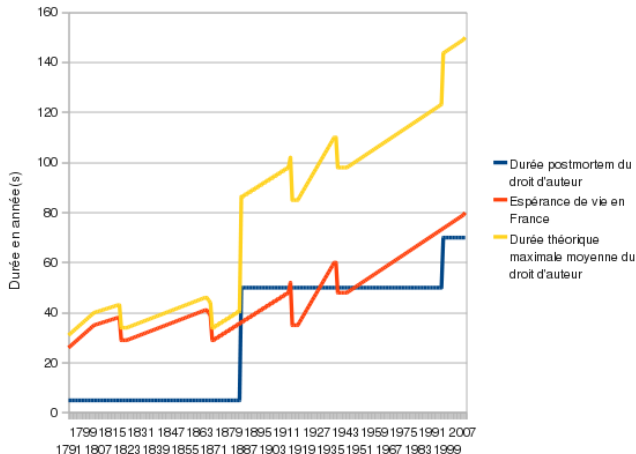
Protection des œuvres limitée dans le temps

- ▶ Directive européenne sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information (1993)
- ▶ traduite dans le droit français par une loi du 27 mars 1997

70 ans après leur mort pour les auteurs

Droit d'auteur : évolution de la durée pour la France

Évolutions de la durée du droit d'auteur
et de l'espérance de vie en France de 1791 à 2007



Exceptions

Concept de *fair use* aux États-Unis (95 ans de droits) et de *fair dealing* dans les pays du Commonwealth

- ▶ fonction du caractère commercial ou désintéressé de l'usage
- ▶ fins éducatives de recherche

En France (code de la propriété intellectuelle)

- ▶ présentation gratuite et privée d'une œuvre dans le cercle familial
- ▶ copie privée (à usage personnel ou pour les proches)
- ▶ courte citation

Alors ?

The Hitchhiker's Guide to the Galaxy

- ▶ Douglas Adams
- ▶ anglais
- ▶ mort en 2001

Les Misérables

- ▶ Victor Hugo
- ▶ français
- ▶ mort en 1885

Alors ?

The Hitchhiker's Guide to the Galaxy

- ▶ Douglas Adams
- ▶ anglais
- ▶ mort en 2001



Les Misérables

- ▶ Victor Hugo
- ▶ français
- ▶ mort en 1885

Alors ?

The Hitchhiker's Guide to the Galaxy

- ▶ Douglas Adams
- ▶ anglais
- ▶ mort en 2001



Les Misérables

- ▶ Victor Hugo
- ▶ français
- ▶ mort en 1885



Alors ?

The Hitchhiker's Guide to the Galaxy

- ▶ Douglas Adams
- ▶ anglais
- ▶ mort en 2001



Les Misérables

- ▶ Victor Hugo
- ▶ français
- ▶ mort en 1885



dépend de l'édition...

Ensuite : domaine public (*copyleft*)



Toutes les œuvres proposées en téléchargement sur le site du Projet Gutenberg sont tombées dans le domaine public

<http://www.gutenberg.org/>

Quid de ?

images trouvées sur le Web

Quid de ?

images trouvées sur le Web

Vérifier les droits

Quid de?

blogs

Quid de?

blogs



Quid de?

Doctissimo

Quid de?

Doctissimo



Quid de?

Twitter

Quid de ?

Twitter

?

« œuvre de l'esprit » ?

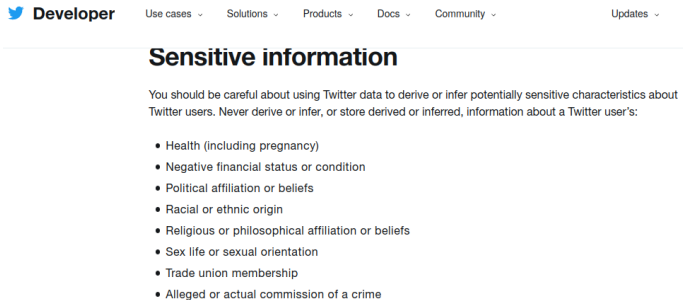
- ▶ originalité
- ▶ mise en forme

<https://scinfolex.com/2015/07/30/twitter-le-micro-plagiat-et-la-physique-quantique-du-copyright/>

Quid de ?

Twitter

Nouvelles règles d'utilisation de Twitter :



The screenshot shows the Twitter Developer website. The top navigation bar includes the Twitter logo, the word 'Developer', and several menu items: 'Use cases', 'Solutions', 'Products', 'Docs', 'Community', 'Updates', and a help icon. Below the navigation bar, the main heading is 'Sensitive information'. The text below this heading states: 'You should be careful about using Twitter data to derive or infer potentially sensitive characteristics about Twitter users. Never derive or infer, or store derived or inferred, information about a Twitter user's:'. A bulleted list follows, detailing the types of sensitive information that should not be derived or inferred from Twitter data.

Developer Use cases Solutions Products Docs Community Updates :

Sensitive information

You should be careful about using Twitter data to derive or infer potentially sensitive characteristics about Twitter users. Never derive or infer, or store derived or inferred, information about a Twitter user's:

- Health (including pregnancy)
- Negative financial status or condition
- Political affiliation or beliefs
- Racial or ethnic origin
- Religious or philosophical affiliation or beliefs
- Sex life or sexual orientation
- Trade union membership
- Alleged or actual commission of a crime

<https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases>

Quid de?

Facebook

?

Facebook

Date: Tue, 27 Oct 2015 22:31:12 -0700
From: "Eric Ringger" <ringger@cs.byu.edu>
Subject: [Corpora-List] Facebook's policy with regard to sharing content
To: "'Manuel Burghardt'" <manuelburghardt@gmx.de>,<corpora@uib.no>

Greetings, Manuel.

Thanks for checking regarding the Facebook dataset you have gathered.

I followed up with folks here at Facebook and found that the short answer to your question is that publishing Facebook data is essentially not permissible. The written policy for external developers can be found here:

<https://developers.facebook.com/policy>

Two points from the policy regarding the API for data access are worth highlighting: one must both "Obtain consent from people before publishing content on their behalf" and "Protect the information you receive from us against unauthorized access, use, or disclosure."

As I'm sure you understand, user privacy is a top priority for everyone at Facebook. I also know the team is open to feedback from the research community regarding these policies. I'd be happy to pass along any feedback you may have.

Regards,

--Eric

Research Scientist, Facebook
Associate Professor, CS, BYU

Introduction

Corpus et corpus

Corpus et droits

Nul n'est censé ignorer la loi

Licences libres

Pour finir

Principes hérités du logiciel libre

« Gnu's Not Unix »

« les utilisateurs ont la liberté d'exécuter, copier, distribuer, étudier, modifier et améliorer »

<http://www.gnu.org/philosophy/free-sw.html>

Exemples de licences pour les corpus

Héritées des logiciels

- ▶ LGPL-LR : héritée de LGPL utilisable dans des applications/ressources non libres (L)
- ▶ CeCILL : licence française (CEA CNRS INRIA Logiciel Libre)

Spécifiques aux œuvres et documents :

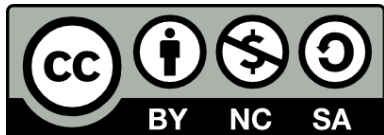
Creative Commons : famille de licences



Licences CC : exemple



Licences CC : exemple



Creative Commons :

- ▶ by : attribution
- ▶ nc : *non commercial* (pas d'utilisation commerciale)
- ▶ sa : *share alike* (partage dans les mêmes conditions) → virale

<https://creativecommons.org/licenses/by-nc-sa/3.0/fr/>

Introduction

Corpus et corpus

Corpus et droits

Pour finir

CQFR : Ce Qu'il Faut Retenir






- ▶ Définition d'un corpus
- ▶ Différentes vues sur les corpus
- ▶ Licences Creative Commons

À faire (2)

Monsieur K. a créé un lexique du macédonien qu'il veut mettre à disposition des chercheurs dans cette langue.

Il voudrait également pouvoir aider un ami qui crée une entreprise de TAL en Macédoine.

Quelle licence lui conseillez-vous ? Pourquoi ?

-  Baude, O. (2007).
Contribution des corpus oraux à la linguistique de corpus : une démarche réflexive
intégrée.
In Actes de Journées de Linguistique de Corpus, Lorient, France.
-  McEnery, T. and Wilson, A. (1996).
Corpus linguistics.
Edinburgh University Press.
-  Sinclair, J. (1996).
Preliminary recommendations on corpus typology.
Technical report, Eagles.