Corpus
exemples, définition, caractérisation

Béatrice Daille

LS2N

Les premiers corpus

- Brown corpus
- Lancaster-Oslo-Bergen (LOB) corpus
- London-Lund Corpus
- COBUILD (Bank of English)
- British National Corpus (BNC)

BNC 1994

- corpus multi-usage de l'anglais moderne 1990
  - 1,5 Giga
  - 100 M de mots
  - 4 049 textes Unicode annotés XML
- langue écrite
  - articles de journaux, articles scientifiques, extraits de livre, essais
  - 3 209 textes (90 % du corpus)
- langue parlée 10 M de mots

Corpus (Sinclair 1996)

« a collection of pieces of language that are selected and ordered according to explicit linguistics in order to be used as a sample of the language »

Données

objet concret – étude qualitative/quantitative

(ling) ensemble limité d'énoncés sur lesquels se base l'étude d'un phénomène linguistique

(lexicométrie) ensemble de textes réunis à des fins de comparaison, servant de base à une étude quantitative (Lebart et Salem 1988)

Linguistique de corpus

corpus linguistics (~ 1980)

étude au moyen d'outils informatiques de corpus constitués de textes informatisés

– textes réels, données attestées
– linguistique descriptive : observation pour reconstituer a *posteriori* des régularités
– construction du corpus

Référentiel actif

- cadre et référentiel de l'analyse
  - éléments à étudier
  - environnement descriptif
- application ou pratique
  - lexicographie
  - terminographie

Sélection des données

- corpus : regroupement raisonné de textes
- archive de textes : regroupement opportuniste

Propriétés d'un corpus (Bommier-Pincemin 1999)

- **Avant** : conditions de signifiance
  - pertinence
  - cohérence
- **Pendant** : conditions d'acceptabilité
  - représentativité (Sinclair 1996)
  - régularité (Sinclair 1996)
  - complétude
- **Après** : conditions d'exploitabilité
  - homogénéité
  - volume

Pertinence

objet d'analyse précis

Règle de pertinence

« Les documents retenus doivent être adéquats comme source d'information pour correspondre à l'objectif qui suscite l'analyse » (Bardin 1977)

Cohérence

le corpus comme entité autonome

Règle de cohérence

« Les documents retenus doivent être homogènes, c'est-à-dire obéir à des critères de choix précis et ne pas présenter trop de singularité en dehors des critères de choix » (Bardin 1977)

Représentativité

Définir un échantillon représentatif

Règle de représentativité

« On peut, lorsque le matériel s'y prête, effectuer l’analyse d'un échantillon. L'échantillonnage est dit rigoureux si l'échantillon est une partie représentative de l'univers de départ. Dans ce cas les résultats obtenus sur échantillon seront généralisables à tout ensemble »(Bardin 1977)

Régularité

Expliciter les principes pour définir le corpus, sans se permettre d'exception

Règle de l'exhaustivité

« une fois défini le champ du corpus, il faut prendre en compte tous les éléments de celui-ci » (Bardin 1977)

Régularité

Explicitation nécessaire

Corpus CoMéRé

https://hdl.handle.net/11403/comere

<post xml:id = "cmr-slr-c001-a2860“ when= "2008-05-01T09:49:36“ who = "#cmr-slr- c001-p000424" type = "sms">

Oui ver20h mc do st benoit vu ke mi mange la ba. tu mange avan de venir? Tu me sone kan t la?

<reg type = "transortho“ > <seg xml:lang = "fra“ cert = "medium“ > Oui vers 20h Mac Do Saint Benoît vu que </seg>
<seg xml:lang = "cpf“ >mi manj la ba. </seg>
<seg xml:lang = "fra“ > Tu manges avant de venir ? tu me sonnes quand t’es là ?</seg>
<add type ="F"><seg xml:lang = "cpf“ cert = "low“ > Wi vèr 20h Mac Do Sin Benoi vu ke</seg> </add>
<add type = "trad"> <seg xml:lang = "fra“ > je mange là-bas </seg> </add> </reg>
</post>

Complétude

Avoir un niveau de détail adapté aux besoins de l'analyse

Règle de l’exhaustivité

« adéquation du modèle à construire à la totalité des ses éléments implicitement contenus dans le corpus » (Greimas 1966)

Homogénéité

Sachant l'objectif de l'analyse et les dimensions de variations que l'on veut étudier, le corpus doit être aussi homogène que possible pour ses autres caractéristiques

Règle de l'homogénéité

« toutes les grandeurs recensées sont des quantités de même nature » (Bardin 1977)

Volume

- permettre des analyses statistiques
- faire émerger des aspects caractéristiques et informatifs

Nombreux « petits » corpus

Corpus Dede http://www.cnrtl.fr/corpus/dede

Corpus (Habert 2000)

« un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et **extralinguistiques** explicites pour servir d'échantillon déterminés **d'une langue** »

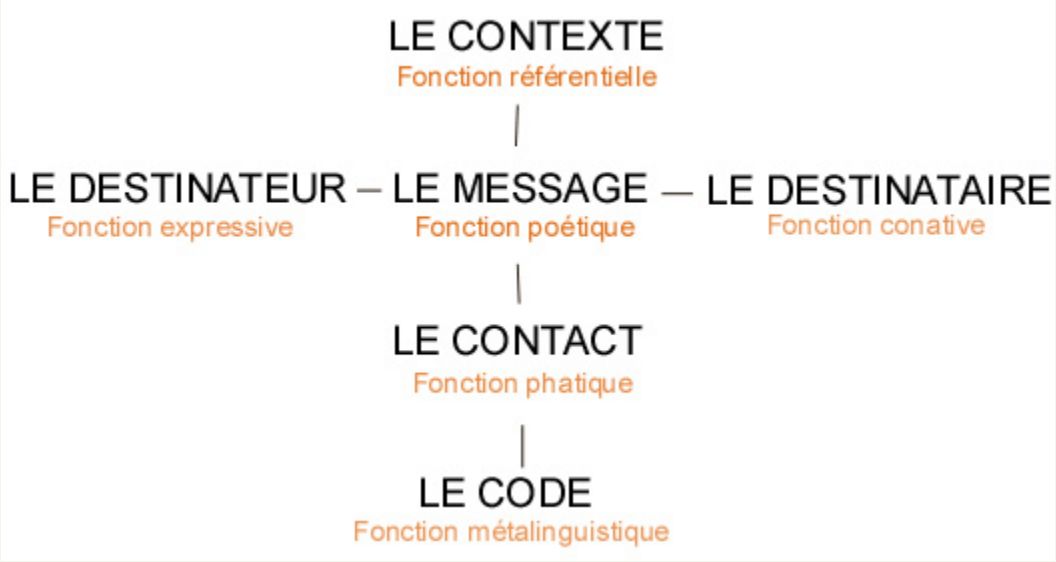
Critères extralinguistiques

- Situation de communication
  - Caractériser le destinateur et le destinataire
  - Fonction du texte : didactique, normatif, descriptif
- Rattachement à des catégories préétablies
  - genres et registres
  - sujets et domaines

  Exemple : Sinclair EAGLES 1996

Modéle de communication

Jakobson 1963



Genres

Biber 1989

« les genres sont les catégories distinguées spontanément par les locuteurs confirmés d'une langue ; par exemple les genres de l'anglais incluent les romans, les articles de journaux, les éditoriaux, les articles de recherches, les discours en public, les nouvelles radiophoniques et la conversation de tous les jours. »

Registres

Biber 1995

Conception élargie des genres

Catégories intuitives qu'utilisent les locuteurs pour répartir les productions langagières

Évolution d’une culture à l’autre et au fil du temps

Genres et registres

"social type of communicative actions, characterized by a socially recognized communicative purpose and common aspect of form."

(Crowston and Williams, 2000)

Un genre caractérise des documents similaires par leur contenu (thème et sujet) et leur forme (traits physique et linguistiques observables)

"a property which distinguishes a class of texts that answers to certain practical interests, and which is associated with a characteristic set of computable structural or linguistic properties."

(Kessler et al. 1997)

public visé, extension à de nouveaux genres

Genre

Définition

<contenu, forme, fonction>

1. Un contenu textuel
2. Des caractéristiques identifiables
3. Une intention communicative

Le genre devrait être prédictible automatiquement

Intentions de communication de l'énonciateur (Sharoff 2010)

- ***Information*** : textes purement informatifs
- ***Discussion*** : textes incluant des prise de positions et des opinions
- ***Instruction*** : textes éducatifs
- ***Inventaire*** : liste d'informations structurées
- ***Loi*** : contrats, lois
- ***Promotion*** : textes dont la finalité est la vente d'un produit, d'un service ou pour convaincre politiquement
- ***Reportage*** : textes factuels relatant des évènements

Typologie a priori/Typologie a posteriori

- A priori : catégories pré-établies
  - Démarche déductive (approche descendante)
- A posteriori : catégories émergent des textes
  - Utilisation de traitements statistiques
  - Démarche inductive (approche montante)
  - Caractérisation des textes par un ensemble de traits linguistiques **Dimensions** (Biber 1989)

Des registres aux dimensions (1/2)

- Registres : emplois du langage définis situationnellement et fonctionnellement
- Objectif : faire émerger des constellations de traits linguistiques ou *dimensions* mobilisées diversement selon les registres
- Démarche : étiquetage de 67 traits linguistiques (temps, aspect, pronom, passifs, modaux, etc.)
- Corpus : 1 000 premiers mots de 481 textes d'anglais contemporain écrit et oral des corpus LOB et London-Lund

Des registres aux dimensions (2/2)

- Mise en évidence de constellations de traits : certains traits ont tendance à apparaître ensemble, et d'autres traits non
  - Orientation narrative
    - Verbes au passé, pronoms 3ème personne, verbes publics (complain), propositions participiales
  - Orientation non-narrative
    - noms, mots longs, prépositions, adverbes de lieu
- Un registre se caractérise par la mobilisation +/- forte de telle ou telle dimension (fiction/ orientation narrative), par la cooccurrence variable des traits
- Méthode : statistique multidimensionnelle (Lexico 3)

5 dimensions

1. Production impliquée ↔ production informationnelle
2. Orientation narrative ↔ orientation non-narrative
3. Référence explicite ↔ référence dépendant de la situation d'énonciation
4. Visée persuasive explicite
5. Style abstrait

8 genres de textes

Obtenus par classification à partir des dimensions

1. Interaction interpersonnelle intime
2. Interaction informationnelle
3. Exposé scientifique
4. Exposé savant
5. Fiction narrative
6. Récit
7. Reportage situé
8. Argumentation impliquée

Genres du web

Plusieurs typologies : de la plus concise (8 genres) à la plus exhaustive (2 000 genres)

(Crowston et al. 2010)

Genres du web

Documentation du web consultée par un informaticien

23 genres et 10 sous-genres (Montesi et Navarette 2008)

Pages à propos	Publicités
Articles - articles scientifiques - articles journalistiques	Documentation -documentation scientifique -documentation législative
Blogs	Brochure
Téléchargements	FAQ
Manuels (5 sous-genres)	Pages d'accueil
Page d'index	Pages à propos
Archives de listes de diffusion	Autres
Présentations	Pages de produits
Pages de ressources	Comparaisons
Page de résultats de recherche	Fil de discussion
Image et video (1 sous-genre page écran)	Wikis et entrées encyclopédiques

Domaines

Eagles (1996)

- Typologie des textes
  http://www.ilc.cnr.it/EAGLES96/texttyp/node36.html%23SECTION00011200000000000000
- Typologie des sujets
  http://www.ilc.cnr.it/EAGLES96/texttyp/node37.html%23SECTION00011300000000000000

domaines génériques : domaine des sciences et technologies

sous-domaines : pour les sciences et technologies : informatique, physique, …

thèmes : pour informatique : intelligence artificielle, science des données, …

Trésor de la langue française (TLF): 758 domaines organisés hiérarchiquement avec 3 niveaux

Cambridge International Dictionary of English (CIDE): 900 domaines organisés hiérarchiquement avec 4 niveaux

Autres traits

Niveau de langue : populaire/familier/courant/soutenu

Variétés de langue standard : français métropole, québécois, français antilles

Niveau de spécialisation

<u>document scientifique</u>

destinateur : spécialiste Destinataire : spécialiste

<u>document vulgarisé</u>

destinateur : spécialiste Destinataire : grand public

destinateur : grand public Destinataire : grand public

Caractérisation Destinateur/Destinataire

Caractérisation du destinateur

- un auteur
- plusieurs co-auteurs
- collectif

Du texte, des textes

Pery-Woodley 1995

du texte : seul matériau linguistique

des textes : unité texte, prise en compte du discours, des conditions de leur production et de leurs objectifs

Corpus prototypique (Gries et Berez 2017)

- Consists of one or more *machine-readable* Unicode text files;
- Is meant to be *representative* for a particular kind of speaker, register, variety or language as a whole, which meant that the sampling scheme of the corpus represents the variability of the population it is meant to represent;
- Is meant to be *balanced*, which meant that the sizes of the subsamples (of speakers, registers, varieties) are proportional to the proportions of such speakers, registers, varieties, etc. in the population the corpus is meant to represent,
- Contains data from *natural communicative settings*, which meant that at the time the language data in the corpus were produced, they were not produced solely for the purpose of being entered into a corpus, and/or the production of the language data was as untainted by the collection of those data as possible.

Les corpus nationaux

- BNC (1993)
- ANC American national corpus
- COCA : Corpus of Contemporary American English

  monitor corpus 1990-2012 (100 M)
- The Mannheim German Reference Corpus (DeReKo)
- Russian National Corpus
- The National Corpus of Polish

Portails corpus

Ortholang https://www.cnrtl.fr/

CLARIN https://www.clarin.eu/portal

ELRA http://catalogue.elra.info

LDC https://www.ldc.upenn.edu/

Corpus et ressources

ANR « CORPUS ET OUTILS DE LA RECHERCHE EN SCIENCES HUMAINES ET SOCIALES » (2007-2010)

Corpus Scientext

Un corpus d'écrits scientifiques variés et des outils logiciels permettant d'effectuer une étude linguistique du positionnement et du raisonnement dans les écrits scientifiques, à travers des raisonnement dans les écrits scientifiques, à travers des marques linguistiques.

Public visé : Linguistes, épistémologues, spécialistes de la recherche d'information, scientifiques

Corpus oraux français

- Corpus de référence du français parlé

  134 enregistrements, 36 heures, 440 000 mots
- TCOF : Traitement de Corpus Oraux en Français

  517 transcriptions /125 H
- PFC (Phonologie du Français Contemporain)

  Accents du français 400 locuteurs

Les corpus pour le TAL

Campagnes d’évaluations

SEMEVAL

https://aclanthology.coli.uni-saarland.de/venues/semeval

CLEF

https://dblp.org/db/conf/clef/index

DEFT

https://deft.limsi.fr/FT

Références

D. Biber. *Representativeness in corpus design*. Cambridge 1993

Alexander Mehler, Serge Sharoff, Marina Santini G*enres on the Web*. Editors Text, Speech and language Technology volume 42 Springer 2010

John Sinclair *Corpus, concordance, collocation*. Oxford. 1991

Mise en œuvre

1. Sur le web

Caractérisation en fonction de l'intension communicative les pages web de la liste d'URLs fournie. Justifier.

1. http://www.emagister.fr/formation_licence_energie_eolienne_photovoltaique-ec2382108.htm
2. http://seme.cer.free.fr/index.php?cat=eoliennes
3. http://gaiamag.over-blog.com/article-livre-les-guerres-du-climat-harald-welzer-38923738.html

Mise en œuvre

2. Avec NLTK

Chapitre 2 : Section 2.1 et 2.2 - Accessing Text Corpora and Lexical resources.
Natural Language Processing with Python. O'Reilly 2011.

Fonctionnalités basiques (Tableau 2.3)

Calcul de fréquences et fréquences conditionnelles (Tableau 2.4)

https://www.nltk.org/book/ch02.html

Examiner les catégories des textes fournis pour chaque corpus ci-dessous et calculer la distribution par catégorie

a) corpus Brown

b) corpus Reuters

c) corpus TIMIT

Et commenter.