Construction et exploitation de corpus

Annie.Tartier

Annie.Tartier@univ-nantes.fr

Master FLE : Université de Nantes

1. Corpus : notions générales
2. Étapes de construction d'un corpus
3. Corpus enrichis
4. Corpus multilingues
5. Outils d'exploitation de corpus

Ce cours s'appuie sur un certain nombre d'ouvrages, et en particulier sur [Bowker L. & Pearson J. 2002]

1. **Corpus : notions générales**
2. Étapes de construction d’un corpus
3. Corpus enrichis
4. Corpus multilingues
5. Outils d’exploitation de corpus

1.1 Linguistique de corpus

- méthodes ***empiriques*** pour étudier l'usage des langues
- basées sur ***attestation*** et non sur ***intuition***

➔ grande quantité de ***matériau textuel***

➔ nécessité de ***moyens informatiques***

1.2 Définition d'un corpus

Définition de John Sinclair (1996) :

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

- grande collection de textes ***authentiques***
- mémorisés sous forme ***électronique***
- rassemblés selon un ensemble de ***critères spécifiques***
- avec un ***objectif*** d'étude précis

► attention à l'usage abusif du mot corpus

1.3 Exemples

Voir document annexe

1.4 Typologie des corpus (1)

- ***corpus de référence*** : représentatif d'une langue donnée dans son ensemble
  - écrit et parlé
  - mélange de genres : journaux, rapports, radio, télé, débats, etc.
  - différents thèmes : vie quotidienne des locuteurs
- ***corpus spécialisé***
  - thème particulier
  - genre particulier
  - communauté de locuteurs
  - langue de spécialité

1.4 Typologie des corpus (2)

- ***corpus écrits***
- ***corpus oraux*** : transcriptions écrites de paroles prononcées (discours, débats, etc.)

- ***corpus monolingues***
- ***corpus multilingues***
  - *corpus parallèles* : textes écrits dans ***une*** *langue d'origine* et leurs *traductions* dans d'autres langues
  - *corpus comparables* : textes écrits dans ***leurs*** *langues d'origine*, ayant des *points communs* (thème, époque, etc.)

1.4 Typologie des corpus (3)

- ***corpus synchronique*** : textes écrits dans la langue d'une même époque
- ***corpus diachronique*** : textes d'époques différentes
- ***corpus fermé*** : construit une fois pour toutes (archive)
- ***corpus de suivi*** (monitor corpus) : reçoit des ajouts réguliers au cours du temps
- ***corpus d'apprenants***

1.5 Utilité des corpus (1)

- ***limites des dictionnaires*** :
  - incomplétude
  - peu d'information contextuelle
  - mise à jour
    - absence de nouveaux termes
    - présence de termes obsolètes
    - lenteur de mise à jour
- ***limite des textes imprimés :***
  - lecture intégrale impossible
  - beaucoup de temps avant de trouver l'information pertinente
- ***limite des experts :*** difficiles à trouver
- ***limite de l'intuition :*** partialité

1.5 Utilité des corpus (1)

intérêt des ressources électroniques :

- beaucoup de matériau en peu de place
- information contextuelle
- mise à jour rapide
- outils de consultation rapides
- observation de l'attesté

1.6 Caractéristiques des corpus

- taille : un petit corpus spécialisé vaut mieux qu'un gros corpus général
  ordres de grandeur : de x 1000 à x 100000 pour les langues de spécialité
- extraits de textes ou textes entiers
- structure : nombre de textes, taille des textes, nombre d'auteurs différents (20000 mots = 20 * 1000 textes de 20 auteurs différents ou 2 longs textes du même auteur)
- écrit / oral retranscrit
- thème
- genres des textes
- source
- langue
- période

1. Corpus : notions générales
2. **Étapes de construction d'un corpus**
3. Corpus enrichis
4. Corpus multilingues
5. Outils d'exploitation de corpus

2.1 Acquisition des droits

- des compromis inévitables
- un corpus imparfait peut être utile à condition de connaître ses défauts et d'en tenir compte au moment de l'interprétation des résultats

1) ***copyright*** et ***autorisations*** : textes électroniques soumis au copyright comme les textes imprimés

➔

- contacter ***auteurs*** et ***éditeurs***
- expliquer les ***objectifs***
- demander les ***autorisations*** (contrat d'utilisation)

2.2 Matières premières

- ***web*** :
  - origines *non contrôlables*
  - dépendance des *moteurs de recherche*
  - matériau le plus souvent *multimedia*
  - *hypertexte* ➡ récupération difficile
- ***CDROM*** : (journaux, encyclopédies, etc.)
  - données plus fiables
  - pas d'accès direct aux textes (logiciels d'exploitation propriétaires)
- ***textes imprimés*** :
  - *numérisation* puis *reconnaissance de caractères*
- ***parole*** :
  - *retranscription manuelle* ou *reconnaissance vocale*
- ***bases de données textuelles***

2.3 Sélection des textes

- élaborer les ***critères de choix***
  - en fonction des *objectifs* de l'étude
  - en respectant les *critères de qualité* (taille, représentativité, etc.)
- ***paradoxe*** (instrument de mesure destructif)
  - sélectionner est *indispensable* pour disposer d'un corpus homogène et représentatif
  - mais sélectionner *agit et modifie* les données observées
- éventuellement constitution *raisonnée* des fragments
- étape ***la plus délicate*** de la construction de corpus

2.4 0rganisation physique du corpus

- ***encodage*** des caractères (isolatin, UTF8, projet UNICODE)
- ***format*** des textes (source et intégrés au corps)
  - texte *brut* (txt)
  - texte *avec balises* (SGML, HTML, XML, . . .)
  - texte *compilé* (pdf, ps, doc, . . .)
  - . . .
- ***structure*** du corpus :
  - un texte (ou un fragment) par fichier
  - plusieurs textes (ou fragments) par fichier

2.5 Annotation du corpus

Cette étape, facultative, fait l'objet d'un chapitre spécial

2.6 Documentation du corpus

- ***indispensable*** pour qu'il soit utilisable
- ***description complète*** du contenu et de la structure
- manuel de ***maintenance*** (corpus ouvert)
- manuel d'***utilisation***

1. Corpus : notions générales
2. Étapes de construction d’un corpus
3. **Corpus enrichis**
4. Corpus multilingues
5. Outils d’exploitation de corpus

3.1 Définition

corpus enrichis, annotés, étiquetés

- corpus au sein duquel ont été intégrées des ***annotations*** ou ***marques***
- historiquement : ***annotations*** ou ***marques*** destinées aux typographes pour agir sur la présentation du texte
- actuellement moyen de marquer les effets de style dans un traitement de texte
- deux types de textes entrelacés :
  - *texte informatif* = contenu du corpus
  - *information sur le texte* = meta information portée par les annotations
  - être capable de les distinguer et ou de les séparer à tout moment

3.2 Méthodes d'annotation

- différentes manières de mettre des annotations
- type de marquage directement lié au logiciel d'exploitation
- difficile de communiquer simplement entre différents systèmes de marquage
- ➔ outils permettant la standardisation de l'annotation des corpus :
  - 1960 : SGML (Standard Generalized Markup Language)
  - maintenant (1998) XML (eXtended Markup Language)
  - projet CES : Corpus Encoding Standard
  - TEI : Text Encoding Initiative

3.3 Nature des annotations

trois catégories d'information :

- ***documentation*** : meta données (langue, taille, . . .)
- ***composition du texte*** :
  - titres, sections, corps de texte, notes
  - paragraphes
  - phrases
  - mots
- ***informations linguistiques*** :
  - *parties du discours* : catégorie grammaticale, genre, nombre, temps du verbe
  - *annotations syntaxiques* : structure de la phrase
  - *annotations sémantiques* : traits sémantiques

3.4 Outils d'annotation

- ***programmes de segmentation*** utilisent des *délimiteurs*, des *patrons* (expressions régulières) et des *listes d'exclusion* (stop liste)
- ***étiqueteurs grammaticaux***
  - attribution d'étiquettes à partir de lexiques ou de dictionnaires
  - mots inconnus
  - désambiguïsation à l'aide du contexte et/ou de calculs statistiques
- ***lemmatiseurs***
  - utilisent les marques grammaticales
  - calculent les formes canoniques (lemmes)
- ***annotation manuelle***

1. Corpus : notions générales
2. Étapes de construction d’un corpus
3. Corpus enrichis
4. **Corpus multilingues**
5. Outils d’exploitation de corpus

4.1 Corpus parallèles

- textes écrits dans leur ***langue d'origine*** et leurs ***traductions*** dans une ou plusieurs autres langues
- deux textes « parallèles » ne sont pas forcément la traduction l'un de l'autre, mais peuvent être la traduction, dans deux langues différentes, d'un même troisième
- on ne sait pas toujours quel est le texte d'origine
- il peut y avoir des traductions de traductions (communauté européenne)
- ***alignement*** pour exploiter ces corpus

4.2 Sources pour corpus parallèles

- organismes spécialisés : LDC, ELRA, . . .
- textes de la communauté européenne
- mémoires de traduction des entreprises de traduction
- revues ayant des correspondances (Scientific American, Pour la science, Bild der Wissenschaft)

4.2 Préparation à l'alignement

Préparation ***manuelle*** ou réalisée par des ***programmes***

- numéroter les paragraphes
- supprimer les retours chariots inutiles
- si un paragraphe de A correspond à deux paragraphes de B, concaténer les deux paragraphes de B en insérant un symbole spécial de manière à pouvoir restituer B dans sa forme originelle
- si un paragraphe de A n'est pas traduit dans B, insérer un paragraphe fictif « paragraphe non traduit »
- si un paragraphe existe dans B qui ne correspond à rien dans A, ajouter dans A un paragraphe fictif « paragraphe ajouté dans la traduction » (dans quelles langues)

4.3 Programme d'alignement

- création de liens entre les paragraphes et les titres qui se correspondent
- création de liens entre les phrases qui se correspondent
- calcul d'un score pour chaque paire de phrase mises en correspondance
- correspondance n'est pas forcément biunivoque.

4.4 Corpus comparables

- ensemble de textes en différentes langues qui ne sont pas les traductions les uns des autres
- textes choisis pour être mis ensemble parce qu'ils ont un certain nombre de caractères communs : thème, type de texte, période
- pas de correspondance entre des parties de textes comme c'est possible dans un corpus comparable

1. Corpus : notions générales
2. Étapes de construction d’un corpus
3. Corpus enrichis
4. Corpus multilingues
5. **Outils d’exploitation de corpus**

5.1 Analyse statistique

Première étude d'un texte : comptages pour ***chaque texte***
et pour ***tout le corpus***

- nombre de mots
- nombre de (vocables | lexèmes | formes de mots | types)

- longueur des mots
- nombre de mots de chaque longueur
- nombre de phrases

5.2 Listes de mots (1)

Liste des ***mots*** et de la ***fréquence*** de leurs occurrences

classée par :

- ordre alphabétique
- ordre alphabétiques des fin de mots
- groupe de mots (cluster, digrams, trigram)
- fréquence croissante (hapax) ou décroissante (mots grammaticaux en tête)
- en excluant les mots d'une liste d'exclusion (stop liste)

5.2 Listes de mots (2)

Attention :

- les homographes sont confondus (nom et verbe)
- les différentes formes grammaticales (singulier / pluriel, formes conjugées) d'un mot sont comptées de manières distinctes
- les mots sont sortis de leur contexte
- traitement des mots composés ou des locutions dépend de la segmentation

➡ recherche de *mots clés* : ceux qui ont une fréquence anormalement élevée dans un texte par rapport à leur fréquence dans d'autres textes

5.3 Concordanciers

- visualisent l'usage des mots dans leur *contexte*
- format KWIC (keyword in context)
- *concordance bilingues* si corpus alignés

- on peut faire varier la longueur des contextes gauche et droit
- par défaut les lignes de concordances sont dans l'ordre d'apparition dans le texte
- on peut classer les lignes de concordances
- on peut filtrer les lignes de concordances
- *expressions régulières* pour obtenir des concordances plus élaborées (un verbe et toutes ses formes)

Éléments de bibliographie

Références

[Biber D. 1994] BIBER, D. (1994) : “Representativeness in corpus design”, *Linguistica Computazionale*, vol. IX-X, pp. 377–408.

[Bowker L. & Pearson J. 2002] BOWKER, Lynne ; PEARSON, Jennifer (2002) : *Working with Specialized Language : a practical guide to using corpora*, New York, Routledge.

[Habert B. *et al.* 1997] HABERT, Benoît ; NAZARENKO, Adeline ; SALEM, André (1997) : *Les linguistiques de corpus*, Paris, Armand Colin / Masson.

[Habert B. *et al.* 1998] HABERT, Benoît ; FABRE, Cécile ; ISAAC, Fabrice (1998) : *De l’écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*, Paris, InterEditions.

[Sinclair J. 1995] SINCLAIR, John (1995) : *Corpus Concordance Collocation*, Oxford University Press.