Institut européen
des métiers de la **traduction** | IEMT
Université de Strasbourg

Web, corpus, traduction : exploitations
Introduction

Enzo Doyen

2025 - M1

🕮 Plan

Informations organisationnelles

- **12** heures de cours dans l'Amphi Pangloss de **18 h à 19h30** ;
  - du **22 septembre** au **10 novembre**.
- **examen** : dossier maison (plus d'informations ultérieurement).

Informations organisationnelles

Page Moodle

Page « Web, corpus et traduction » sur Moodle :

https://moodle.unistra.fr/enrol/index.php?id=10364

Contact

En cas de problèmes, questions :

enzo.doyen@unistra.fr

Objectifs du cours

- comprendre pourquoi et comment utiliser les corpus dans une perspective traductionnelle ;
- apprendre à créer ses propres corpus ;
- apprendre à manipuler les outils d'analyse et d'exploitation des corpus.

En traduction, les corpus sont partout

Linguee

Reverso Context

MemoQ

Antidote

I. Définition d'un « corpus »

Qu'est-ce qu'un « corpus » ?

En premier lieu, quels **critères principaux** utiliser pour définir un corpus ?

Qu'est-ce qu'un « corpus » ?

En premier lieu, quels **critères principaux** utiliser pour définir un corpus ?

1. le type/support ;
2. la langue et le type de langue ;
3. la taille ;
4. l'objectif.

Définition d'un corpus : type/support

- corpus textuels ;
  - papier ;
  - numérisations ;
  - publications en ligne.
- corpus oraux ;
  - enregistrements ;
  - transcriptions.
- corpus multimodaux.

Définition d'un corpus : langue et type de langue

Un corpus peut être monolingue ou multilingue.

Pour les corpus multilingues, on distingue les corpus **parallèles** des corpus **comparables**.

Définition d'un corpus : langue et type de langue
Corpus parallèle ou comparable ?

Définition :

Corpus parallèle : rassemble des documents originaux et leurs traductions directes.

Corpus comparable : rassemble des documents originaux dans différentes langues. Les documents ne sont pas des traductions directes, mais traitent du même sujet.

Définition d'un corpus : langue et type de langue
Corpus parallèle ou comparable ?

EN	FR
While P2P systems had previously been used in many application domains, the concept was popularized by file sharing systems such as the music-sharing application Napster. [...]	La particularité des architectures pair-à-pair réside dans le fait que les échanges se font directement entre deux ordinateurs connectés au système, sans transiter par un serveur central. [...]

Définition d'un corpus : langue et type de langue
Corpus parallèle ou comparable ?

EN	AR
Meetings normally should be held during regular meeting hours, namely, from 10 a.m. to 1 p.m. and from 3 p.m. to 6 p.m., on working days	تُعقد الاجتماعات عادة خلال ساعات الاجتماعات العادية، أي من الساعة 00/10 إلى الساعة 00/13 ومن الساعة 00/15 إلى الساعة 00/18، أيام العمل؛

Déf. d'un corpus : langue et type de langue

Notion d'alignement

EN	FR
Steam Deck OLED has 30-50% more battery life. We fit a bigger battery into the case, and the OLED display draws less power. Combined with the updated, more efficient AMD APU, you have way more time to play your favorites.	L'autonomie de Steam Deck OLED est 30 à 50 % supérieure à celle du modèle LCD , ce grâce à une plus grande batterie et à l'écran OLED, qui est moins énergivore. Ajoutez à cela un nouvel APU d'AMD plus efficace, et vous obtenez encore plus de temps pour jouer à vos jeux favoris.

Déf. d'un corpus : langue et type de langue

Notion d'alignement

EN	FR
<s>Steam Deck OLED has 30-50% more battery life.**</s>** **<s>**We fit a bigger battery into the case, and the OLED display draws less power.**</s>** **<s>**Combined with the updated, more efficient AMD APU, you have way more time to play your favorites.**</s>**	**<s>**L'autonomie de Steam Deck OLED est 30 à 50 % supérieure à celle du modèle LCD**</s>** **<s>**, ce grâce à une plus grande batterie et à l'écran OLED, qui est moins énergivore.**</s>** **<s>**Ajoutez à cela un nouvel APU d'AMD plus efficace, et vous [...].**</s>**

Définition d'un corpus : taille

Problèmes éventuels posés par un corpus **trop petit** ?

Définition d'un corpus : taille

Problèmes éventuels posés par un corpus **trop petit** ?

Manque d'informations, biais de sélection, analyses limitées

Problèmes éventuels posés par un corpus **trop grand** ?

Définition d'un corpus : taille

Problèmes éventuels posés par un corpus **trop petit** ?

Manque d'informations, biais de sélection, analyses limitées

Problèmes éventuels posés par un corpus **trop grand** ?

Qualité des données, problèmes techniques/d'exploitation

Définition d'un corpus : taille

Problèmes éventuels posés par un corpus **trop petit** ?

Manque d'informations, biais de sélection, analyses limitées

Problèmes éventuels posés par un corpus **trop grand** ?

Qualité des données, problèmes techniques/d'exploitation

Un corpus de grande taille sera toujours préférable à un corpus de petite taille, à condition d'assurer une bonne qualité des données, et de disposer des outils adéquats pour l'exploiter.

La taille dépend aussi de l'**objectif** : selon le domaine, les besoins peuvent varier et il peut être plus ou moins difficile d'obtenir des données textuelles.

Définition d'un corpus : objectif

- analyse des phénomènes et de la variation linguistiques ;
- analyse diachronique ;
- analyse statistique ;
- étude d'un domaine de spécialité.

Définition d'un corpus : objectif

- analyse des phénomènes et de la variation linguistiques ;
- analyse diachronique ;
- analyse statistique ;
- étude d'un domaine de spécialité.

Tous ces objectifs peuvent s'avérer pertinents dans une **perspective traductionnelle**.

Définition d'un corpus

« A corpus is 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. »

Sinclair [1996]

Définition d'un corpus

« [A corpus is] a collection of texts assumed to be representative of a given language, dialect, or other subset of language, to be used for linguistic analysis »

Francis [1992 : 7]

Définition d'un corpus

On peut résumer un « corpus » à un **ensemble de faits de langue**, dans **une ou plusieurs langues**, sélectionnés sur des **critères précis**, dans un **objectif particulier**.

II. Usage des corpus en traduction

Usage des corpus en traduction

Quels sont, selon vous, les potentiels usages des corpus en traduction ?

Usage des corpus en traduction

Objectif principal : amélioration de la **fluidité** et de l'**authenticité** des traductions.

- vérification de l'usage linguistique ;
- recherche documentaire spécialisée ;
- création de glossaires spécialisés ;
- analyse des connotations, des registres de langue et des nuances de sens ;
- aide à la rédaction.

Usage des corpus en traduction : vérification linguistique

« les X dernières minutes » ou « les dernières X minutes » ?

ce et la beauté des	**dernières 20 minutes**	dont le tournage en s
lendemain pour les	**dernières 30 mn**	de cuisson).</s><s>»
puis voit, dans les	**dernières 30 minutes**	, sans rien pouvoir fai
iins à s'exécuter, la	**dernière 60 mins**	.</s><s>J'ai mise en

nerfs dans ces	**10 dernières minutes**	, avec un chassé
passe dans les	**15 dernières minutes**	... Qui sait si l'on
<s>Pendant les	**10 dernières minutes**	: Laisser les élève
teurs actifs des	**5 dernières minutes**	) Le nombre maxi
oyager mais les	**30 dernières minutes**	allaient être d'un

395 résultats pour « les dernières X minutes »

contre 8 725 pour « les X dernières minutes »

Usage des corpus en traduction : vérification linguistique

« encodage par GPU » ou « encodage par le GPU » ?

ajoute l'	**encodage accéléré par le GPU**	pour certains
activer l'	**encodage par le GPU**	.</s><s>Mais
utiliser l'	**encodage par GPU**	avec OBS, p
souci d'	**encodage par le GPU**	puisque ça s
elle de l'	**encodage par GPU**	...</s><s>2 v

9 résultats pour « encodage par le GPU »

contre 3 pour « encodage par GPU »

Usage des corpus en traduction : recherche spécialisée

« inside-out tracking »

t utilizes inside-out or outside-in tracking.</s><s>	**Inside-out tracking**	, typically found in standalone headsets, utilizes built-in sensors, eliminating th
inside-out or outside-in tracking systems.</s><s>	**Inside-out tracking**	uses built-in sensors and cameras in the headset, providing greater freedom o
body and hands.</s><s>The Quest 3 is built with	**inside-out tracking**	, which means the headset uses cameras to track lights that the controllers em
d Reality Headset</s><s>HP Reverb G2</s><s>	**Inside-out tracking**	means less hassle when setting up this headset with a combined resolution of
r needed to run at these resolutions.</s><s>With	**inside-out tracking**	, you also won't need to set up external cameras or base stations, which mean
The Meta Quest 3, HP Reverb G2, and more use	**inside-out tracking**	, which means the headset tracks your movement using cameras and sensors
ing in a smaller space, or are a more casual user,	**inside-out tracking**	might be for you.</s><s>If you want the best experience and have access to e
ets to get up and running on this list.</s><s>With	**Inside-Out tracking**	and hand tracking built-in, you can go from unboxing to up-and-running in VR i
ations.</s><s>The Quest 2 and Vive Cosmos use	**inside-out tracking**	–that is, sensors on the headset instead of placed around your room.</s><s>A

Usage des corpus en traduction : recherche spécialisée

« pastoral accompaniment »

ers and of Catholic healthcare institutions 10.</s><s> **Pastoral accompaniment** and the support of the sacraments 11.</s><s>Pastoral discernment

patient's care, and should be comprised of adequate **pastoral accompaniment** .</s><s>Adequate support must be provided to the families who bea

and human coexistence rooted in justice. 10.</s><s> **Pastoral accompaniment** and the support of the sacraments Death is a decisive moment in th

/ human situation of desolation or discomfort.</s><s> **Pastoral accompaniment** involves the exercise of the human and Christian virtues of empathy

Usage des corpus en traduction : connotations

élève 3,224,796× étudiant 2,545,893×

modifiers of "élève/étudiant"			
mauvais	13,350	178	•••
enseignant	3,008	279	•••
volontaire	3,066	694	•••
ancien	62,715	14,394	•••
studieux	1,494	370	•••
étudiant	5,514	2,050	•••
brillant	4,753	3,526	•••
jeune	17,484	50,145	•••
boursier	1,394	5,089	•••
étranger	2,044	29,512	•••
prêt	218	9,886	•••
universitaire	60	7,546	•••

noun modifiers of "élève/étudiant"			
allophone	1,097	0	•••
officier	936	0	•••
avocat	441	0	•••
dys	288	0	•••
ingénieur	3,059	263	•••
surdoué	349	57	•••
stagiaire	165	320	•••
entrepreneur	45	215	•••
doctorant	15	69	•••
étranger	16	170	•••
locataire	0	80	•••
juriste	0	58	•••

nouns modified by noun "élève/étudiant"			
officier	77	0	•••
fiche	341	0	•••
base	191	0	•••
ancien	629	46	•••
ratio	154	45	•••
effectif	56	23	•••
bail	0	190	•••
syndicat	0	474	•••
mouvement	0	1,480	•••
visa	0	531	•••
logement	0	807	•••
job	0	804	•••

Usage des corpus en traduction : connotations

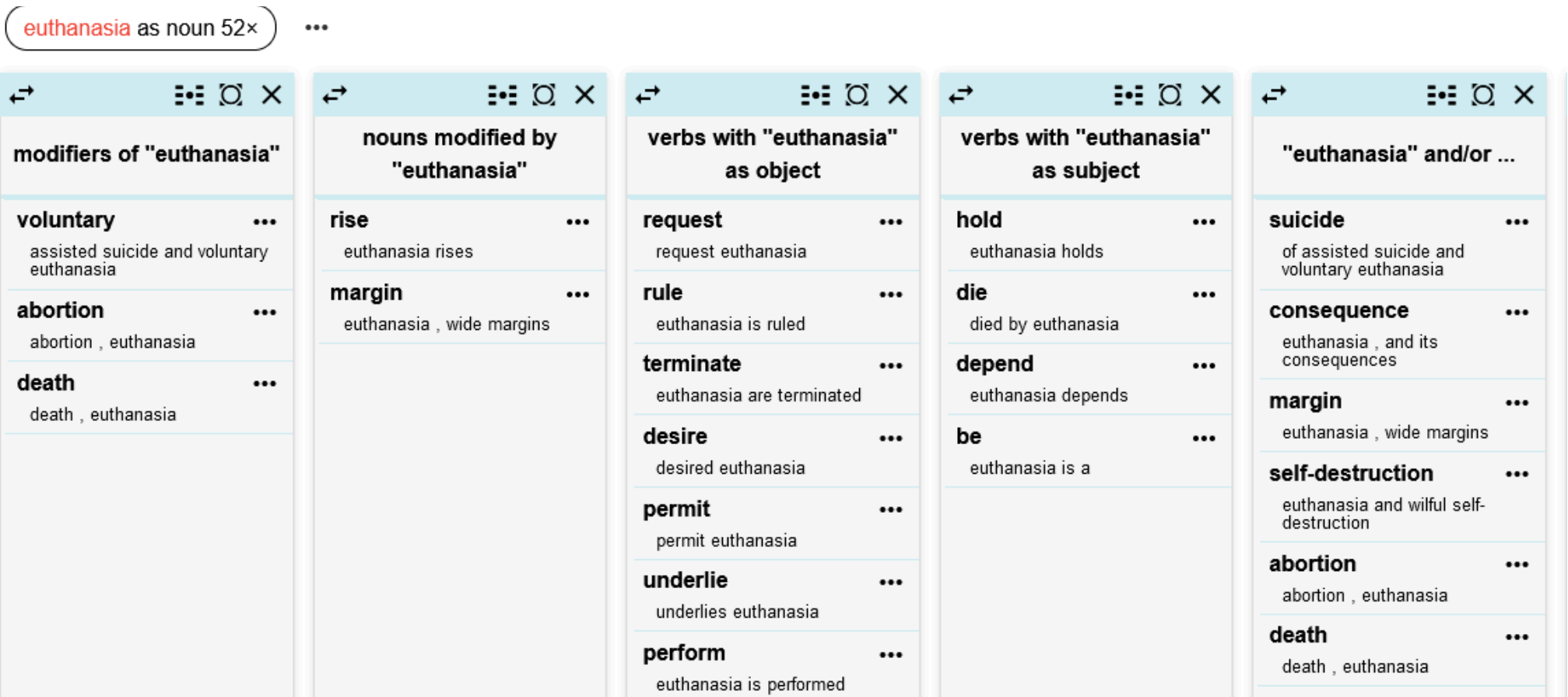
élève 3,224,796× étudiant 2,545,893×

"élève/étudiant" de+les			
lycée	2,203	130	•••
classe	7,935	756	•••
école	12,145	2,811	•••
collège	1,770	375	•••
établissement	1,869	598	•••
section	523	139	•••
filière	443	772	•••
cycle	407	1,856	•••
faculté	42	479	•••
université	126	2,002	•••
campus	15	211	•••
masters	0	184	•••

"élève/étudiant" dans			
degré	288	0	•••
classe	3,639	288	•••
lycée	522	94	•••
collège	462	82	•••
apprentissage	1,714	413	•••
école	2,523	798	•••
besoin	203	382	•••
cursus	100	182	•••
filière	160	379	•••
université	44	838	•••
campus	0	106	•••
faculté	0	156	•••

"élève/étudiant" de			
CM2	5,808	0	•••
maternelle	6,469	42	•••
primaire	5,649	61	•••
terminale	7,069	127	•••
classe	26,219	1,851	•••
école	38,183	9,485	•••
École	6,999	2,757	•••
cycle	3,042	7,439	•••
université	647	9,387	•••
Université	358	9,868	•••
master	115	2,660	•••
Master	119	2,779	•••

Usage des corpus en traduction : aide à la rédaction



Antidote fonctionne aussi !

soin apporté	attention
soin particulier	special care
soins gratuits	free care
soins appropriés	proper care
et 76 autres...	
Avec adjectif classificateur (53)	
soins palliatifs	palliative care
soins intensifs	intensive care
soins infirmiers	medical care
soins médicaux	medical care
soins dentaires	dental care
et 48 autres...	
Avec nom complément (74)	
soins de santé	health care
soins à domicile	home care
soins de longue durée	long-term care
soins de qualité	quality care
soins aux patients	patient care
et 69 autres...	

Antidote fonctionne aussi !

Cornebille pensait que la marquise enfin allait connaître ses soins, depuis plus de dix ans prodigués vainement.

René Boylesve, *la Leçon d'amour dans un parc*, Gallica

A partir de cet instant, la sollicitude la plus éclairée, les soins les plus habiles ne cessèrent de m'être prodigués.

Amédée Delorme, *Journal d'un sous-officier, 1870*, Projet Gutenberg

Les experts seront chargés de vérifier la propreté, la qualité des repas et des soins prodigués aux résidants.

Radio-Canada.ca

Les soins pharmaceutiques sont prodigués aux patients atteints de maladies respiratoires (clientèle ambulatoire et hospitalisée).

Université de Montréal

Intégration des corpus aux outils de TAO

MemoQ : LiveDocs

- prise en charge des documents monolingues/multilingues ;
- alignement automatique des corpus parallèles avec **LiveAlign** ;
- conversion des documents bilingues en mémoires de traduction (MT) avec **ActiveTM** ;
- possibilité de stocker les documents directement dans MemoQ.

Documentation : **https://docs.memoq.com/current/en/Concepts/concepts-livedocs.html**

III. Outils de création, d'analyse et d'exploitation des corpus

Outils de création, d'analyse et d'exploitation des corpus

Sketch Engine

AntConc

Outils de création, d'analyse et d'exploitation des corpus

Sketch Engine (Kilgarriff et al., 2014)

- service en ligne ;
- plus de **800 corpus précréés** disponibles dans **100 langues** différentes ;
- permet de créer facilement des **corpus personnalisés** ;
- outils de **recherche avancée** ;
- accès **gratuit** (corpus de 1 million de mots) pour les universitaires partenaires (dont l'Unistra).

Outils de création, d'analyse et d'exploitation des corpus

AntConc (Anthony, 2024)

- logiciel **gratuit** compatible avec Windows, Mac et Linux ;
- fonctionne **en local** (pas besoin de connexion Internet, pas de stockage des données en ligne) ;
- outils de **visualisation** intégrés ;
- quelques **outils de recherche et d'analyse plus avancés** en comparaison avec Sketch Engine.

IV. Avantages et limites des corpus dans une perspective traductionnelle

Avantages de l'utilisation des corpus

- analyses linguistique et d'usage de la langue sur la base de faits de langue authentiques (possibilité de confirmer des intuitions et de les **étayer par des exemples concrets**) ;
- aide utile lors de la traduction de **textes spécialisés**, notamment en cas de manque d'expertise dans le domaine ;
- **conversion en MT** possible avec les outils adéquats ;
- peuvent être utilisés comme source d'inspiration pour la **rédaction**.

Limites de l'utilisation des corpus

- composition **équilibrée et cohérente** du corpus nécessaire ;
- limitations en fonction du **domaine** ;
- limitations en fonction **des langues de travail** (langues peu dotées).

V. Le Web comme corpus ?

Le Web comme corpus ?

Pensez-vous que le Web puisse être considéré comme un corpus ?

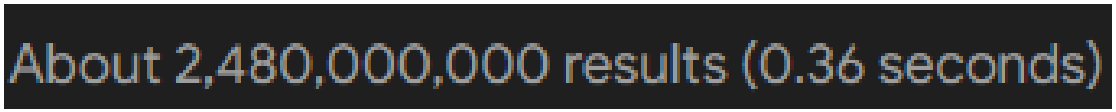
Le Web comme corpus ? (Non !)

« The low-entry-cost way to use the Web is via a commercial search engine. If the goal is to find frequencies or probabilities for some phenomenon of interest, we can use the hit count given in the search engine's hits page to make an estimate. People have been doing this for some time now. [...] **But if the work is to proceed beyond the anecdotal, a range of issues must be addressed.** »

— Kilgarriff (2007)

Le Web comme corpus ? (Non !)

- **fonctionnalités de recherche avancées limitées** en comparaison avec les outils spécialisés ;
- problèmes de **fiabilité** et de **qualité** des données (d'autant plus à l'heure de l'IA générative !).
- types de textes **trop divers** et **non filtrables** ;
- le nombre de résultats correspond au nombre de pages indexées, pas au **nombre d'occurrences des termes**.

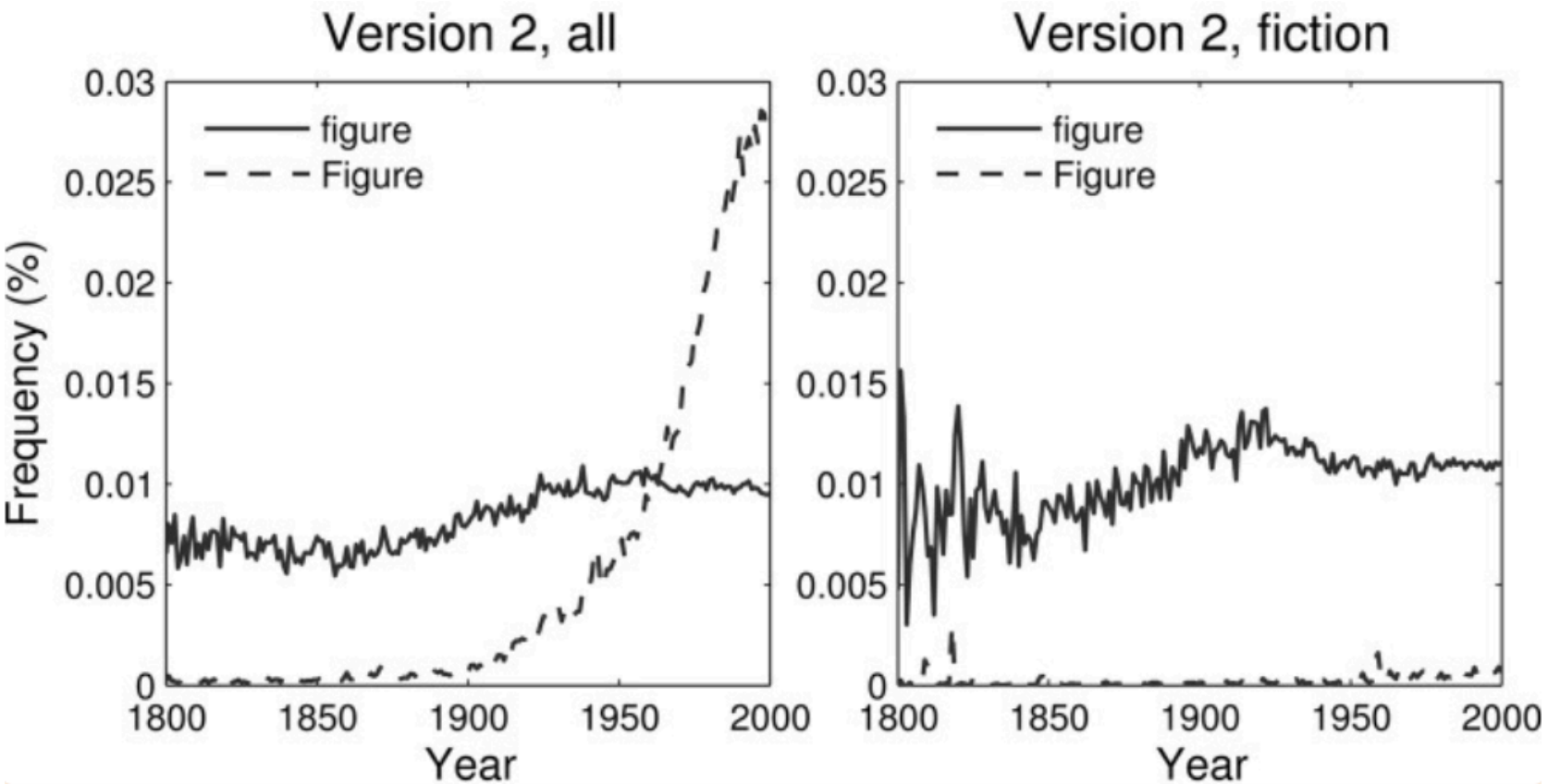


Quid de Google Books ?

C'est déjà mieux (fonctionnalités de recherche plus complètes), mais il reste des limitations :

- **métadonnées et contenu textuel tronqués** pour des raisons de droit d'auteur ;
- **risque de manipulation** des données par l'ajout de livres numériques via Google Play Books ;
- **biais** au niveau des types de textes (notamment, articles scientifiques).

Google Books : biais au niveau des types de textes



Pechenick et al. (2015)

Google Books : biais au niveau des types de textes

Version modifiée de Google Books proposée par le *Corpus of Historical American English* (COHA) : **https://www.english-corpora.org/googlebooks/**

- syntaxe de recherche plus avancée ;
- collocations ;
- comparaison des fréquences ;
- …

Le Web comme corpus : résumé

L'utilisation du Web et/ou de Google Books dans une perspective traductionnelle peut être utile pour des **vérifications** ou de la **recherche documentaire rapides**…

Le Web comme corpus : résumé

L'utilisation du Web et/ou de Google Books dans une perspective traductionnelle peut être utile pour des **vérifications** ou de la **recherche documentaire rapides**…

… mais ne saurait remplacer un corpus spécialisé en raison de ses **limitations intrinsèques** (fiabilité, qualité des données, diversité des textes) et **extrinsèques** (fonctionnalités des moteurs de recherche).

VI. Corpus personnalisés

Corpus personnalisés : objectif et portée

Avant de créer un corpus, il est essentiel de définir clairement **son objectif** et **sa portée**.

Corpus personnalisés : objectif et portée

Avant de créer un corpus, il est essentiel de définir clairement **son objectif** et **sa portée**.

- Quel est le domaine concerné ?
- Quels types de textes pourraient être pertinents ? Sont-ils facilement accessibles ?
- Quelle est la visée du corpus (analyses, constitution de glossaires, recueil de documents spécialisés, etc.) ?
- Existe-t-il déjà un ou des corpus pour mon domaine ?

Sélection des textes

Existe-t-il déjà un ou des corpus pour mon domaine ?

Base de textes pour l'étude du rêve

Fiches ajoutées récemment : Murakami, Bauby

Récits littéraires

Récits de rêveurs ordinaires

Interprétation et théorie des rêves

Recherche: par auteur / dans les récits de rêves

« Si l'on réunissait les rêves que les hommes ont eus pendant une période définie, on verrait surgir une image très juste de l'esprit de cette période. »
Hegel.

https://reves.ca/index.php

Sélection des textes

Existe-t-il déjà un ou des corpus pour mon domaine ?

SKETCH ENGINE

https://www.sketchengine.eu/corpora-and-languages/corpus-list/

800 corpus prêts pour utilisation dans 100 langues

Sélection des textes

Existe-t-il déjà un ou des corpus pour mon domaine ?

DOW JONES FACTIVA

Frantext

Factiva (presse) (depuis le site de la BU Unistra : bu.unistra.fr)

Frantext (littéraire) (depuis le site de la BU Unistra : bu.unistra.fr)

HuggingFace https://huggingface.co/datasets

Corpus personnalisés : sélection des textes

Comme vu précédemment, on ne peut pas utiliser le Web comme un corpus…

Corpus personnalisés : sélection des textes

Comme vu précédemment, on ne peut pas utiliser le Web comme un corpus…

… **mais** on peut s'en servir pour **constituer** son propre corpus !

Corpus personnalisés : création de corpus depuis le Web

Deux manières de faire :

- approche **manuelle** ;
- approche **automatisée**.

Corpus personnalisés : création de corpus depuis le Web

Approche manuelle

L'approche **manuelle** consiste à choisir soi-même les textes pertinents.

Corpus personnalisés : création de corpus depuis le Web

Approche manuelle

1. **Recherche directe selon le contenu** : on peut déjà connaitre les textes que l'on veut ajouter à son corpus, ou en trouver à l'aide de moteurs de recherche (Google, Bing, etc.) ;
   - On copie/colle ensuite chaque texte dans un fichier texte.

Corpus personnalisés : création de corpus depuis le Web

Approche manuelle

2. **Utilisation de fichiers** : récupération de fichiers locaux déjà existants (PDF, EPUB, MOBI, fichiers de localisation…).
   - Possibilité également de faire une recherche en ligne sur Google (`file: pdf`, `file: epub`, etc.).

All Images Videos Books Web News Finance Tools

Pdf download | 1st year | Multiplication

UC Davis Math
https://www.math.ucdavis.edu › linear-guest PDF

Linear Algebra

... **Linear Algebra**? . . . . . 367. G.2 Systems of ... **file** the size of the matrix is given, after which each number is a ...
436 pages

Atlantic International University
https://students.aiu.edu › resources › onlineBook PDF

Introduction to Linear Algebra, 4th Edition

This provides video lectures of the full **linear algebra** course 18.06. MATLAB® is a registered

Corpus personnalisés : création de corpus depuis le Web

Approche manuelle

L'ensemble des fichiers récupérés, à la fois par recherche directe et par utilisation de fichiers existants, constitue le corpus.

Corpus personnalisés : création de corpus depuis le Web

Approche manuelle

Avantages : contrôle total et facile sur les textes ajoutés ; pas de problème de compatibilité des formats ; possiblement moins de nettoyage des données nécessaire.

Inconvénients : processus long et fastidieux ; risque d'oublier des textes pertinents.

Corpus personnalisés : création de corpus depuis le Web

Approche automatique

SketchEngine

BootCaT

Corpus personnalisés : création de corpus depuis le Web
Approche automatique

Avantages : processus rapide et automatisé ; possibilité de récupérer un grand nombre de textes en peu de temps.

Inconvénients : risque de récupérer des textes non pertinents (si pas de filtrage) ; peut nécessiter de nettoyer les données ; certains formats non pris en charge[1].

[1] Par exemple, SketchEngine ne prend pas en charge les formats EPUB/MOBI. Mais il y a toujours possibilité de combiner les deux approches (manuelle et automatique) !

Création de corpus depuis le Web : web scraping

Ce que font **SketchEngine** et **BootCaT**, c'est récupérer automatiquement des textes sur le Web : on appelle cela du **web scraping**.

Création de corpus depuis le Web : web scraping

Ce que font **SketchEngine** et **BootCaT**, c'est récupérer automatiquement des textes sur le Web : on appelle cela du **web scraping**.

Possibilité d'utiliser des méthodes alternatives selon les besoins et les connaissances techniques.

Exige d'avoir des connaissances minimales en programmation (HTML/CSS, Python ou JavaScript).

Création de corpus depuis le Web : web scraping

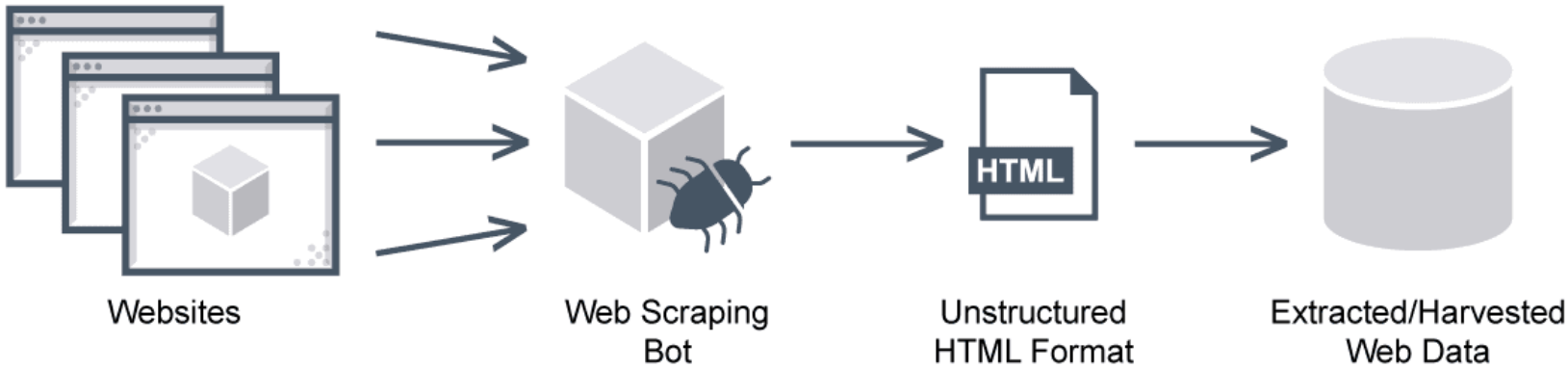


Diagramme de fonctionnement d'un web scraper

Source : **https://avinetworks.com/**

Outils de web scraping

Beautiful Soup (Python)

Selenium (Python, Java, JavaScript, Ruby…)

Modèles de langue (Ahluwalia et Wani, 2024 ; Huang et al., 2024)

Outils de web scraping

Avantage : peut permettre d'obtenir des meilleures performances en comparaison avec les outils généralistes (moins de temps de nettoyage).

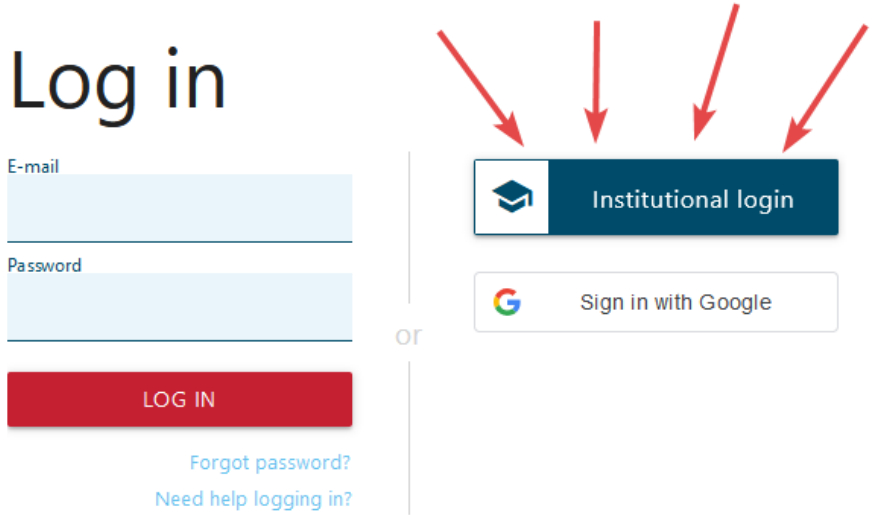
Inconvénient : nécessite des connaissances en programmation.

Conclusion

- en traduction, les corpus sont **partout**, et ils sont utiles autant pour les **outils de traduction** que pour les **traducteurs et traductrices professionnel·les** ;
- il existe une large gamme d'outils destinés à la **création, l'analyse et l'exploitation des corpus** ;
- prendre le temps d'utiliser ces outils et/ou de composer ses propres corpus (selon les besoins) peut grandement **améliorer la qualité et l'authenticité des traductions**.

Pour la semaine prochaine

Jetez un rapide coup d'œil à **Sketch Engine** et vérifiez que vous pouvez vous connecter avec votre compte Unistra.



Bibliographie

Ahluwalia, A., et Wani, S. (juin 2024). *Leveraging Large Language Models for Web Scraping* (Numéro arXiv:2406.08246). arXiv. 10.48550/arXiv.2406.08246

Anthony, L. (2024). *AntConc (Version 4.3.1) [Computer Software]*.

Huang, W., Gu, Z., Peng, C., Li, Z., Liang, J., Xiao, Y., Wen, L., et Chen, Z. (septembre 2024). *AutoScraper: A Progressive Understanding Web Agent for Web Scraper Generation* (Numéro arXiv:2404.12753). arXiv. 10.48550/arXiv.2404.12753

Kilgarriff, A. (2007). Last Words: Googleology Is Bad Science. *Computational Linguistics*, *33*(1), 147‑151. 10.1162/coli.2007.33.1.147

Kilgarriff, A., Baisa, V., Bŭsta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., et Suchomel, V. (2014). The Sketch Engine: Ten Years On. *Lexicography*, *1*, 7‑36.