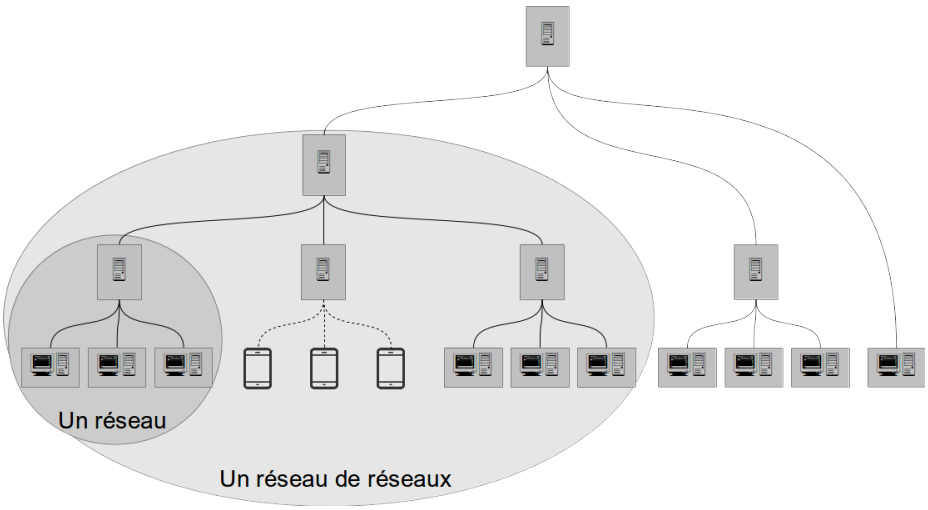
http://pro.aiakide.net, achille.falaise@univ-paris-diderot.fr, loic.liegeois@univ-paris-diderot.fr

Internet pour les SdL : le *Web* comme corpus

A. Petit préalable: qu'est-ce qu'Internet ?

Un réseau informatique (*net* en anglais) est un ensemble d'*ordinateurs* (fixes, portables, smartphones...) connectés à un *routeur* (box Internet, antenne de téléphonie mobile...) par des câbles physiques ou des liaisons radio. Chaque routeur est lui même relié à un ou plusieurs autres routeurs, et ainsi de suite.



Voir par exemple la carte du réseau du fournisseur d'accès à Internet (FAI) Free : https://www.free-reseau.fr/statistiques/

De nombreux réseaux (FAI, entreprises de télécommunication, universités, organismes de recherche, etc.) sont aujourd'hui interconnectés entre eux, formant un vaste *réseau de réseaux*, mondial. Le principal *réseau de réseaux* a été baptisé *Internet* en 1983, mais quelques grand réseaux sensibles (armée, transports, électricité...), sont toujours généralement séparés.

⚠ Internet ≠ Web ≠ Google !

Un peu d'histoire

Après la création du premier modem en 1958, le premier réseau transcontinental (américain) est créé en 1969 (4 ordinateurs). En 1971, le réseau ARPANET est créé avec 23 machines. En 1973 l'Angleterre et la Norvège rejoignent le réseau.

ARPANET en 1974 : https://fr.wikipedia.org/wiki/ARPANET#/media/File:Arpanet_1974.svg

D'autres réseaux existent, p. ex. X.25 en Europe. Le *Minitel*, créé en 1980, se base sur ce réseau.

Les réseaux servent alors surtout pour les *emails*. Les premiers *forums* apparaissent en 1979 (*newsgroups*).

En **1983**, ARPANET fusionne avec d'autres réseaux et est rebaptisé ***Internet***. 1 000 ordinateurs connectés en 1984, 10 000 en 1987, 100 000 en 1989… 368 000 000 en 2000.

À partir du langage HTML (créé en 1989) et du protocole HTTP (créé en 1990), Tim Berners Lee et Robert Cailliau, ingénieurs au CERN, vont donner naissance à un ensemble très populaire de pages liées entre elles par des *hyperliens*, qu'on appellera ***World Wide Web*** (et son contenu *pages Web* ou *sites Web*) à partir de **1991**.

La plus ancienne (13/11/90) page de ce qu'on n'appelle pas encore le *Web* est conservée à l'adresse : http://www.w3.org/History/19921103-hypertext/hypertext/WWW/Link.html

Des logiciels sont créés spécialement pour afficher ces pages: les ***navigateurs Web***, par exemple Mosaic (1993), Netscape (1994), Internet Explorer (1995), Firefox (2004), Google Chrome (2008)...

Les ***moteurs de recherche*** apparaissent peu après le *Web* : W3Catalog (1993), Lycos (1994), AltaVista (1995), Yandex (1997), Google et MSN Search (1998), Baidu et Exalead (2000), Yahoo! Search (2004), DuckDuckGo (2008), Qwant (2013)… Il existe aussi des moteurs spécialisés.

- Comparez *Google Search*, ***Google Scholar*** et *Dmoz.org* . Est-ce le même outil ? La même base de données ?

Le Web s'organise en deux espaces: le **Web de surface** (20% du contenu), qui est indexé par les moteurs de recherche, et le **Web profond** (80% du contenu) qui n'est pas accessible par les moteurs de recherche.

Les citations

L'adresse d'une page Web (URL) se compose de 4 parties : protocole**://**domaine/chemin**?**paramètres

1. le protocole (pour le Web : http ou https),
2. le **nom de domaine** (qui se lit de droite à gauche, et « commence » donc par un code tel quel .fr, .ca, .com, .org, etc.),
3. un chemin (facultatif),
4. divers paramètres (facultatif).

- Isolez le nom de domaine de ces deux sites. **Comment citeriez-vous le site en général ? La page en particulier ?** Voir l'exemple en page 4.

http://www.lemonde.fr/voyage/video/2010/10/21/plongee-dans-le-metro-moscovite_1424766_3546.html

http://www.tlfq.ulaval.ca/AXL/francophonie/HIST_FR_s1_Expansion-romaine.htm

Garder une trace d'une page Web : https://web.archive.org

https://web.archive.org/web/20160815120112/http://www.univ-lyon2.fr/universite/presentation/

B. Faire de la linguistique avec Google Search

Avant tout, le Web est-il à proprement parler un *corpus* ? Contient-il des données verbales, sélectionnées selon quel critère, visant quel but de recherche ? Combien de mots ?

En français, dit-on « hôtel à Paris », ou bien « hôtel en Paris » ? Recherchez sur *Google Search*. Quelle conclusion en tirez-vous ?

Dans la plupart des moteurs de recherche, on peut rechercher une expression exacte en l'encadrant par des guillemets. Le caractère * peut remplacer n'importe quel mot. Le mode avancé permet de restreindre à une langue ou bien a un site Web précis.

Essayez avec « en Bron ». Combien *Google Search* annonce-t-il de résultats ? Essayez d'afficher le dernier résultat. Qu'en déduisez-vous ?

En français, dit-on « voyage aux États-Unis », ou bien « voyage en États-Unis » ou « vacances à États-Unis » ? Recherchez sur *Google Search*. Quelle conclusion en tirez-vous ?

Indice, testez ces pages : http://www.plot-generator.org.uk/story/
http://www.megabambou.com/pmg/index.html

Réessayez en vous limitant aux sites lemonde.fr, ou tripadvisor.fr, pour les expressions « à/en États-Unis », puis « à/en Isère ».

Concrètement, quelle utilisation voyez-vous pour Google Search ? Rechercher des exemples sur le Web ? Sur un site précis ? Récupérer un nombre de résultats ? Déterminer si une expression est attestée ?

Lexicographie avec Google Trends

Google Trends est un outil permettant de visualiser les requêtes effectuées par les utilisateurs de Google Search. Il est intéressant en lexicographie, mais attention, ce n'est pas un outil fait pour cela ! Il ne faut pas perdre de vue qu'il se base sur les *recherches* des utilisateurs, pas sur le contenu des sites Web.

Allez sur le site de Google Trends: http://www.google.fr/trends/explore

- Dans quelles régions françaises utilise-t-on le terme *chocolatine* ? Comparez avec *pain au chocolat*.

- Pour *pain au chocolat*, que se passe-t-il en octobre 2012 ? Que pouvez-vous en déduire ?

- Comparez l'utilisation de *tchat* et de *clavardage* en France et au Canada.

- Étudiez les recherche du mot *grippe* dans le temps. Quand ce mot est-il le plus recherché ? Que se passe-t-il en 2009 ?

Concrètement, Google Trends est-il utilisable pour rechercher des régionalismes ? Une évolution linguistique ?

Linguistique diachronique avec Google Ngram Viewer

Google Ngram Viewer est un outil permettant d'effectuer des recherches diachroniques dans Google Books.

Recherchez « dans » en français entre 1600 et 2000. Quelle échelle est-elle utilisée en abscisse ? En ordonnée ? Comparez avec « en ».

Recherchez « ouïr » en français entre 1600 et 2000. Que constatez-vous globalement ? Autour de 1615 ? Comparez[1] « homme/femme », « sécurité/liberté », « filière bovine », « peuple/population/populations ». Quelles sont les limites de ces résultats ?

Comparez l'évolution de l'emploi des prépositions « à » et « en » avec Martinique, Guadeloupe et Haïti.

Essayez avec « Isère ». Quels biais possibles ?

Corpus parallèles avec Linguee

Linguee est un moteur de recherche pour des textes multilingues, notamment des sites et des documents d'organisations internationales. Il permet de rechercher une expression (ce qu'un dictionnaire ne peut pas faire) et d'afficher une série d'exemples de traduction.

Comment traduiriez-vous « conseil de classe » en anglais ? Quelle est le degré de qualité des traductions sur lesquelles Linguee se base ? (inutile de connaître l'anglais, il y a d'autres indices dans la page…)

Essayez avec « par conséquent », « en avoir le cœur net ». Quelle traduction retenir ? Comparez « sans doute » et « sans aucun doute ». Comment traduiriez-vous « *raining cats and dogs* » (facile) et « *off the charts* » (plus difficile) en français ?

Les outils à retenir : Google Search (mode avancé), Google Scholar, Web Archive, Google Trends, Google Ngram Viewer, Linguee.

1 Guillaume Champeau, *5 graphiques étonnants sur le vocabulaire à travers Google Books*, http://www.numerama.com/magazine/27679-5-graphiques-etonnants-sur-le-vocabulaire-a-travers-google-books.html, publié le 02/12/2013, consulté le 28/02/2017.