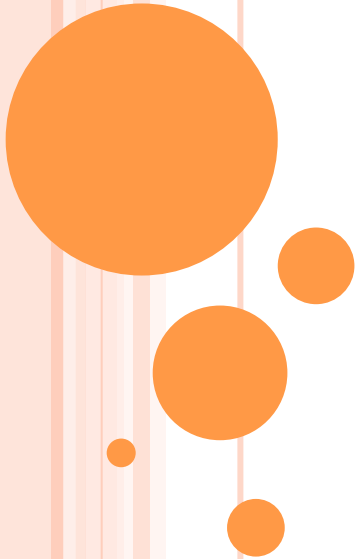


Alignement de chaînes et de textes

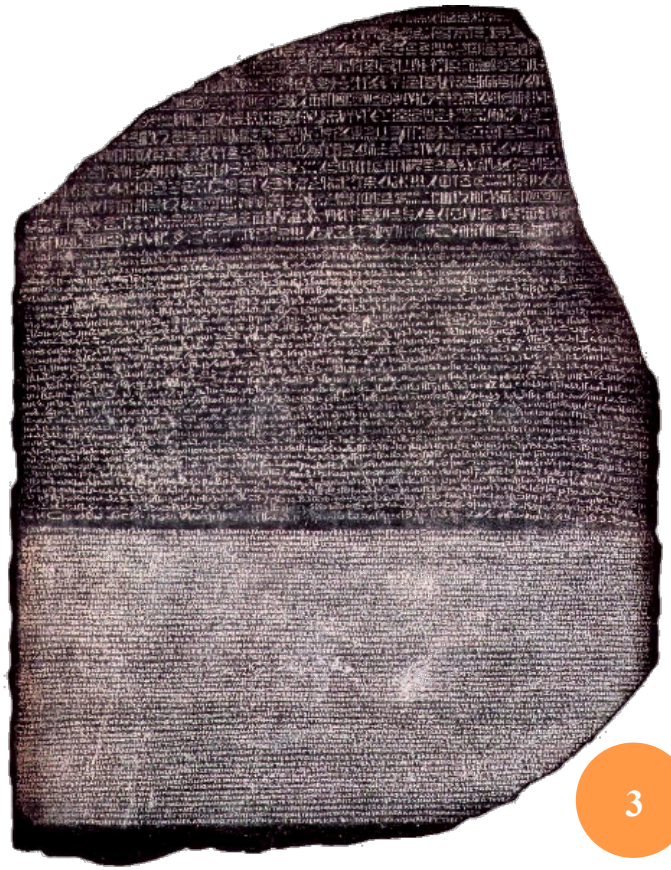
Emmanuel Morin
Master 2 ATAL
2015/2016

I – Motivation

La pierre de Rosette

- La découverte de la pierre de Rosette en 1799 permit à Jean-François Champollion d'apporter en 1822 la clé du déchiffrement de l'écriture hiéroglyphique

- Cette pierre comportait un *texte parallèle* en deux langues (le grec et l'égyptien) et trois écritures (grecque, démotique et hiéroglyphique).



Corpus (1/3)

- Ensemble de documents disponibles sous format électronique

- La constitution d'un corpus nécessite une certaine homogénéité : période, langue parlée ou écrite, genre, type de discours…

CORPUS (2/3)

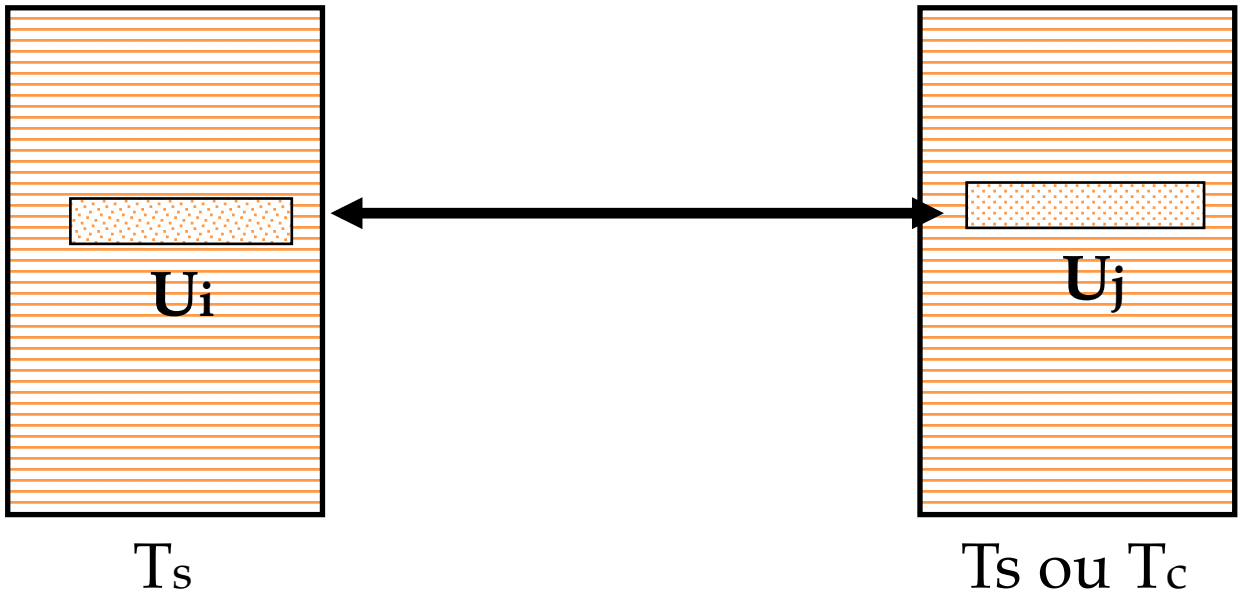
- Deux types de corpus :
  - **corpus monolingues** : corpus dans une langue (Ts pour texte source) : p. ex le BNC un corpus de 100 millions de mots étiquetés représentatif d'une grande variété de situations de communication
  - **corpus bilingues** : corpus composés de deux langues en relation ou non de traduction (Ts et Tc pour textes source et cible) : p. ex le Hansard un corpus de débats du Parlement canadien en anglais et français
    - **parallèles** : un ensemble de textes accompagnés de leur traduction
    - **comparables** : des textes dans des langues différentes qui ne sont pas la traduction l'un de l'autre mais qui abordent une même thématique sur une même période

CORPUS (3/3)

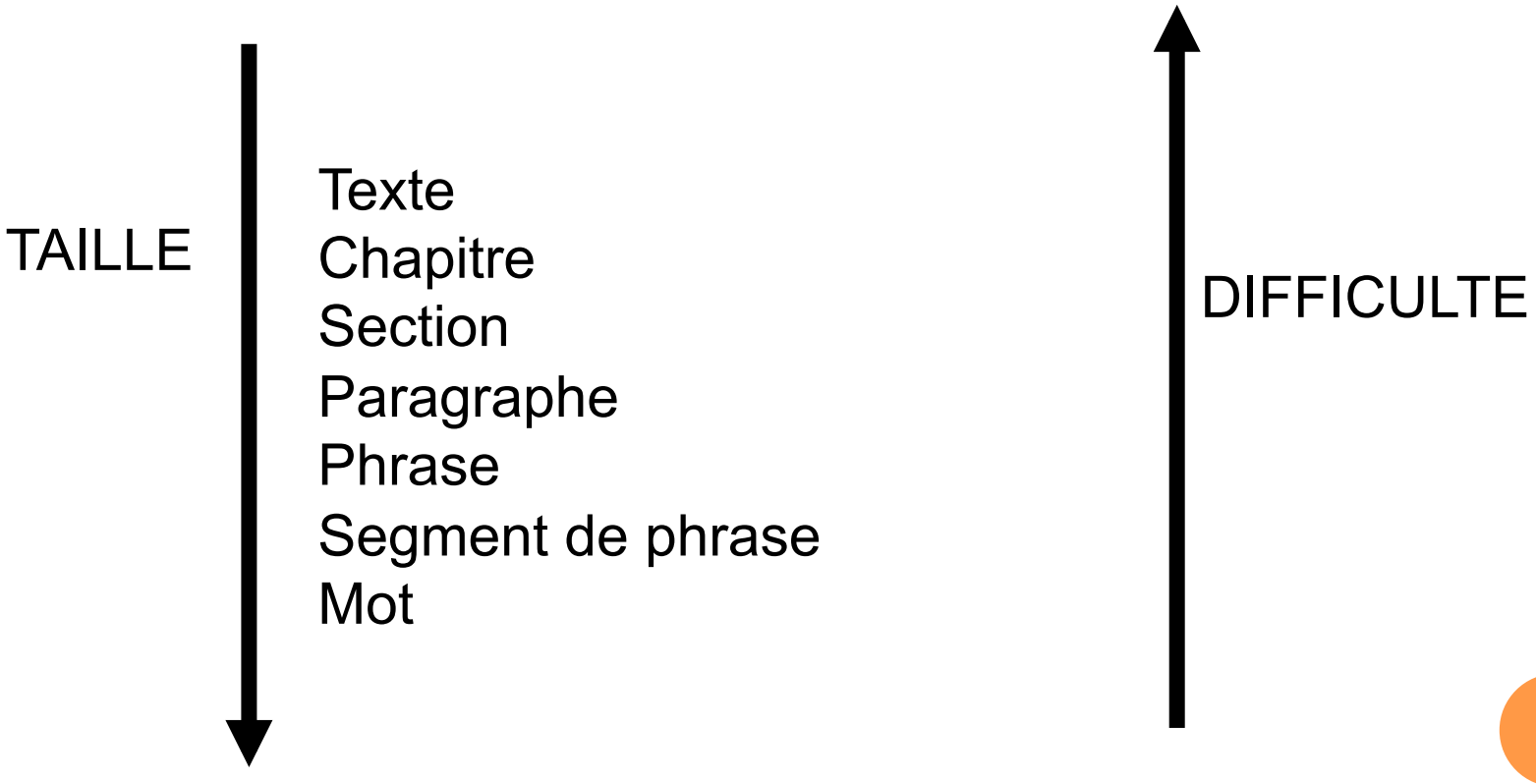
- Ces corpus peuvent être **multimodaux** lorsqu'ils sont constitués de modalités différentes :
  - du texte (écrit sous forme électronique)
  - du texte manuscrit (OCR ou écriture manuscrite)
  - de la parole
  - de la vidéo
  - …

ALIGNEMENT DE TEXTES (1/3)

- Les recherches en alignement de textes s'intéressent à **mettre en correspondance automatiquement des unités** (Ui de Ts avec Uj de Ts ou Tc) **aux sens identiques** à travers l'exploitation de corpus monolingues ou bilingues



ALIGNEMENT DE TEXTES (2/3)



Exemple d'alignement phrastique dans un corpus parallèle

En Europe, l'exploitation à grande échelle des forêts boréales a commencé vers le milieu du XIXe siècle.

Pour alimenter les industries il fallait transporter le bois de plus en plus loin, tandis que les coûts augmentaient aussi.

Or, il n'était pas facile de déplacer les scieries et les communautés avoisinantes, les unes et les autres représentant d'immenses investissements et avantages sociaux.

In Europe, large-scale utilization of boreal forests started around the mid-1800s.

Hauling distances to existing wood processing industries increased, as did the costs, but the mills and surrounding communities could not easily be moved, as they represented huge investments and social benefits.

Exemple d'alignement sous-phrastique dans un corpus monolingue

La ceinture de verdure **protège de façon permanente** des terres précieuses et favorise un environnement plus sain pour l'ensemble des Ontariens et des Ontariennes. → Dans un état de conservation remarquable, la végétation a bénéficié d'une **protection constante**.

Exemple d'alignement de termes complexes et simples dans un corpus comparable

Aujourd'hui, le **dépistage du cancer du sein** par **mammographie** est recommandé, en France, dans toute la population des femmes âgées de 50 à 74 ans.

In North America (especially the USA), **breast screening** is more commonly done on an individual basis rather than being population based, and the decision to do screening **mammography** in patients above age 70 might depend on the clinical situation.

APPLICATION DE L'ALIGNEMENT

- Différents types d'applications sont visées par l'alignement de textes :
  - la traduction automatique :
    - mise en correspondance d'unités textuelles ou phrastiques
  - l'aide à la traduction automatique
    - mise en correspondance d'unités textuelles
  - la recherche d'information monolingue ou multilingue
    - mise en correspondance d'unités sous-phrastiques (en Q/R) ou textuelles (en RI)
    - mise en correspondance d'unités textuelles (en RI multilingue)

II – Indices d'alignement multilingue

INDICES D'ALIGNEMENT MULTILINGUE

- Dans le cadre de l'alignement multilingue, deux indices formels sont privilégiés :
  1. les **transfuges** qui désignent des chaînes de caractères invariantes lors de la traduction : noms propres, nombres, sigles…
  2. les **cognats** qui désignent des équivalents traductionnels présentant une double ressemblance graphique et sémantique (généralement recherche d'un 4-gramme)

(FR) J'ai de sérieuses raisons de croire que la *planète* d'où venait Le Petit *Prince* est l'astéroïde **B 612**.
(EN) I have serious reason to believe that the *planet* from which the little *prince* came is the asteroid known as **B 612**.

IDENTIFICATION DES COGNATS

- Inkpen et al. (2005) distinguent différents niveaux pour les cognats :
  - véritables cognats : *reconnaissance* (FR) et *recognition* (EN)
  - faux amis : *blesser* (EN) et *bless* (EN)
  - cognats partiels dépendant du contexte : *facteur* (FR) et *factor/mailman* (EN)
  - cognats génétiques partageant des origines communes : *chef* (FR) et *head* (EN)
  - mots non reliés : *glace* (FR) et *ice* (EN)

MÉTHODES D'IDENTIFICATION DE COGNATS

- Simard *et al.* (1992) distinguent les candidats cognats suivant trois catégories :
  1. les couples de chaînes identiques comportant au moins un chiffre
  2. les couples de mots commençant par 4 caractères identiques
  3. les signes de ponctuations identiques

- Church (1993) considère tous les couples partageant le même 4-gramme comme des candidats cognats sous réserve que ceux-ci aient une fréquence inférieure à un certain seuil.

MÉTHODES D'IDENTIFICATION DE COGNATS

- Débili and Sammouda (1992) proposent de calculer un indice de ressemblance en fonction des sous chaînes maximales :

$$N = \left(1 - \frac{|l_1 - l_2|}{(l_1 + l_2)}\right) \sum_t n(t) * t^2$$

- Où *l1* et *l2* sont les longueurs de deux à comparer, *n(t)* est le nombre de sous-chaînes communes de longueur *t*

- Cette mesure n'indique pas si deux mots sont des candidats cognats ou non, mais elle permet d'évaluer leur degré de ressemblance

RÉFÉRENCES

- Debili F, Sammouda E. (1992). Appariements de Phrases de Textes bilingues Français-Anglais et Français-Arabes. In *Proceedings of the 14th International Conference on Computational Linguistics*, p. 528-524, Nantes, France.
- Diana Inkpen, Oana Frunză, Grzegorz Kondrak (2005). Automatic Identification of Cognates and False Friends in French and English. In *Proceedings of the 5th International Conference* Recent Advances in Natural Language Processing, p. 251-257, Bulgaria.
- Simard M., Foster G., Isabelle P. (1992). Using cognates to align sentences. In *Proceedings of the 14th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 67-81, Montréal, Canada.

EXERCICE (1/3)

- Données :
  - un corpus comparable français/anglais sur le cancer du sein composé d'articles scientifiques
  - corpus-atal.tgz
    - txt-source/ : les fichiers français prétaités (ne pas tenir compte des fichiers vides) – attention à _ pour les locutions
    - txt-target/ : les fichiers anglais prétaités (ne pas tenir compte des fichiers vides) – attention à _ pour les locutions
    - termer_source/corpus.lem : le corpus français étiqueté et lemmatisé – il y a des marques pour identifier le fichier d'origine – attention aux ambigüités restantes
    - termer_target/corpus.lem : le corpus anglais étiqueté et lemmatisé – il y a des marques pour identifier le fichier d'origine – attention aux ambigüités restantes

EXERCICE (2/3)

- Exemple:
  - fichier français prétaité :
    - Le cancer du sein constitue un problème majeur de santé publique , à_la_fois dans les pays dits développés , où il représente le premier cancer féminin , et dans les pays en voie de développement , où son incidence ne cesse d' augmenter .
  - corpus français étiqueté et lemmatisé:
    - Le/DTN:m:s/le cancer/SBC:_:s/cancer du/DTC:m:s/du sein/SBC:_:s/sein constitue/VCJ:{{1|3}p:s:pst:{ind|subj}| 2p:s:pst:imper}/constituer:1g un/DTN:m:s/un problème/SBC:_:s/problème majeur/ADJ:m:s/majeur de/PREP/de santé/SBC:_:s/santé publique/ADJ:f:s/public ,/, à_la_fois/ADV/à_la_fois dans/PREP/dans les/DTN:_:p/le pays/SBC:m:p/pays dits/ADJ2PAR:m:p/dire…
  - ambigüité restante :
    - {fonde/VCJ:{{1|3}p:s:pst:{ind|subj}|2p:s:pst:imper}/fonder:1g|fonde/VCJ:{1|3}p:s:pst:subj/fondre:3g}

EXERCICE (3/3)

- Exercice
  - Identifier les transfuges et les cognats présents dans le corpus comparable
  - Donner quelques statistiques sur le résultats de votre travail :
    - Combien de transfuges et cognats identifiés ?
    - Quelle est la qualité des transfuges et cognats identifiés ?

POUR MÉMOIRE…

- La qualité peut s'évaluer en termes de :
  - **Précision** : le nombre de réponses pertinentes retrouvées par rapport au nombre de réponses totales proposées par le système
  - **Rappel** : le nombre de réponses pertinentes retrouvées par rapport au nombre de réponses pertinentes présentes dans la ressource à analyser
  - **F-mesure** ou **F-score** : combinaison de précision et rappel

table de contingence	**mots pertinents**	**mots non pertinents**
mots retrouvés	vrais positifs (VP)	faux positifs (FP)
mots non retrouvés	faux négatifs (FN)	vrais négatifs (VN)

$$P = \frac{VP}{VP + FP} \qquad R = \frac{VP}{VP + FN} \qquad F - mesure = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R}$$