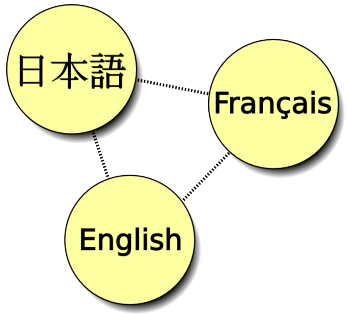
# Alignement multilingue en corpus comparables spécialisés

Emmanuel Prochasson

Laboratoire d'Informatique de Nantes Atlantique
Encadré par Béatrice Daille et Emmanuel Morin

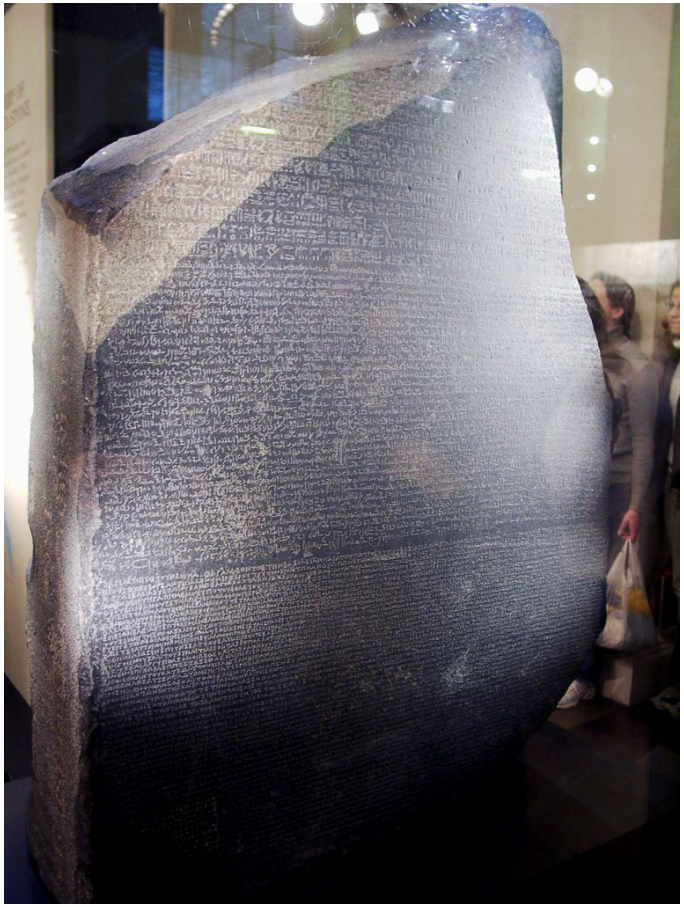



17 décembre 2009

UNIVERSITÉ DE NANTES

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

# Introduction



Reproduction de la Pierre de Rosette,

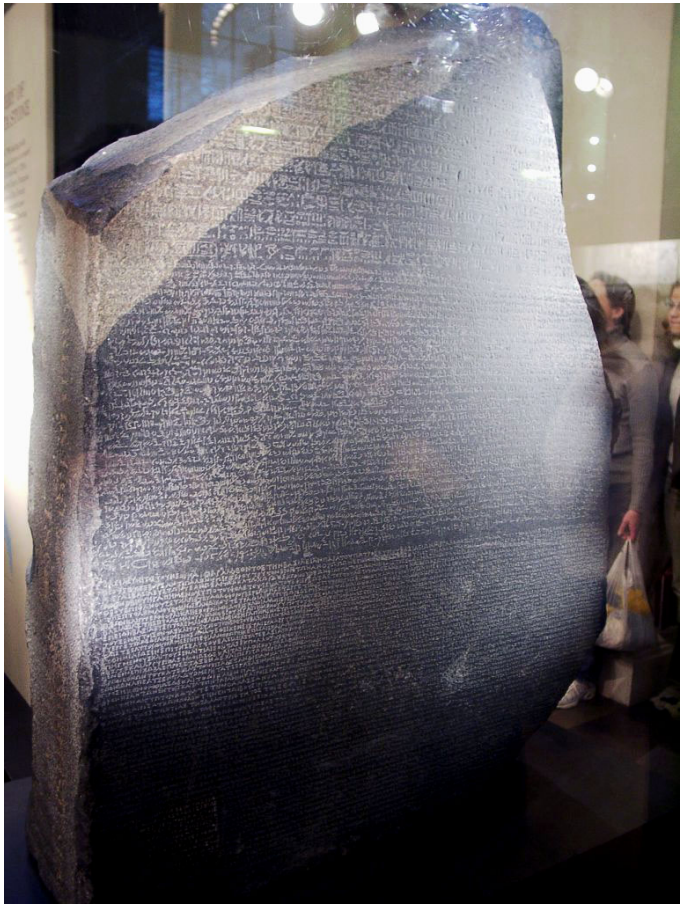
British Museum

La *Pierre de Rosette*

- Trois textes équivalents, deux langues, trois écritures
- Aligné par Champollion (1822)
- Déchiffrage des hiéroglyphes égyptiens

Exemple de *corpus multilingue*

# Introduction



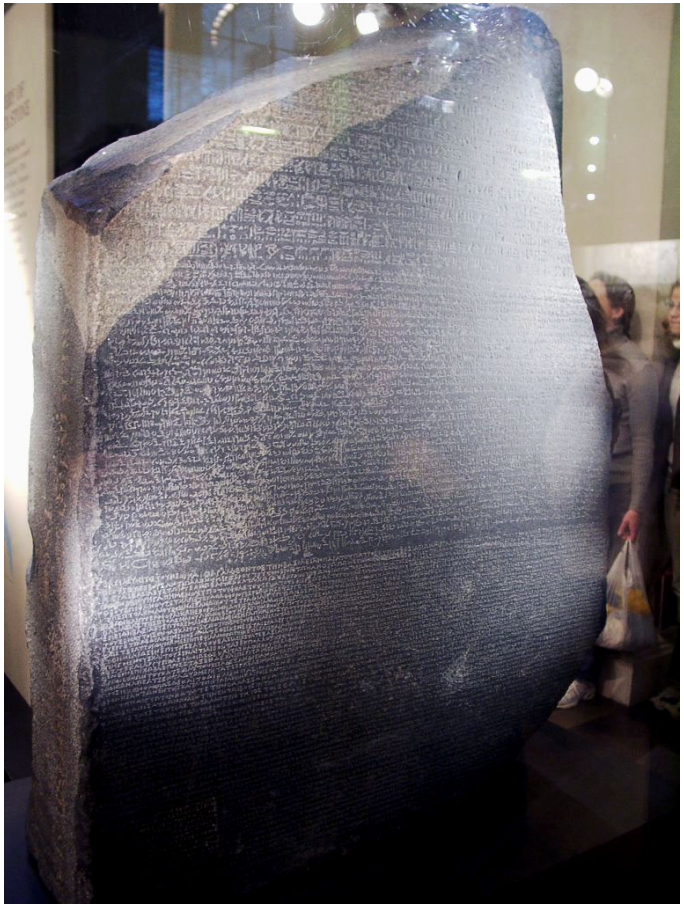
Reproduction de la Pierre de Rosette,

British Museum

*Exploitation des corpus multilingues*

- Étude du processus de traduction

# Introduction



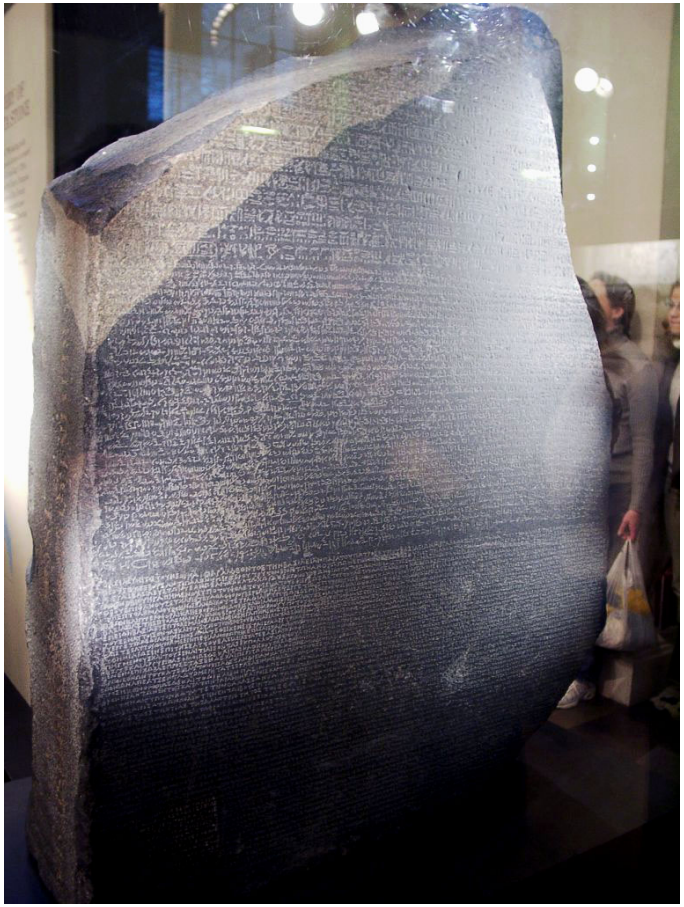
Reproduction de la Pierre de Rosette,

British Museum

*Exploitation des corpus multilingues*

- Étude du processus de traduction
- Mémoire de traduction

# Introduction



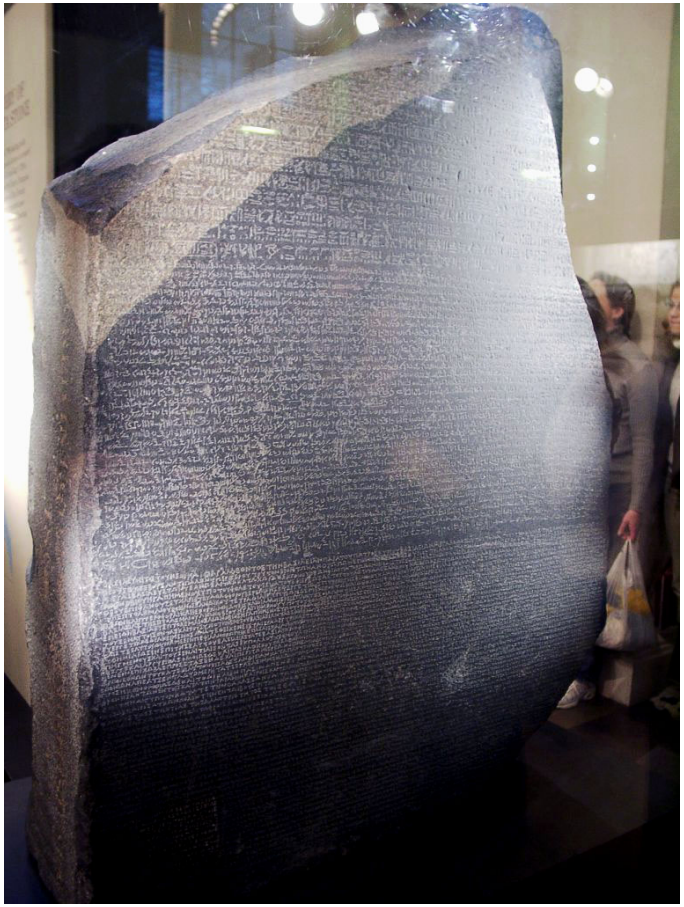
Reproduction de la Pierre de Rosette,

British Museum

*Exploitation des corpus multilingues*

- Étude du processus de traduction
- Mémoire de traduction
- Analyse contrastive multilingue

# Introduction



Reproduction de la Pierre de Rosette,

British Museum

*Exploitation des corpus multilingues*

- Étude du processus de traduction
- Mémoire de traduction
- Analyse contrastive multilingue
- Lexicographie/Terminologie multilingue

# Corpus multilingues

Deux types de corpus multilingues :

- les corpus *parallèles*
- les corpus *comparables*

# Corpus multilingues

Deux types de corpus multilingues :

- les corpus *parallèles*
  - regroupent des documents en relation de traduction
  - coûteux, généralement réservés à certains domaines
- les corpus *comparables*

# Corpus multilingues

Deux types de corpus multilingues :

- les corpus *parallèles*
  - regroupent des documents en relation de traduction
  - coûteux, généralement réservés à certains domaines
- les corpus *comparables*
  - regroupent des documents multilingues non-traductions
  - disponibles dans de nombreuses langues
  - plus difficiles à exploiter

# Corpus comparables

## Bowker & Pearson, 2002

« *Un corpus comparable est composé d'ensembles de textes, dans des langues différentes, qui ne sont pas des traductions mutuelles.* »

# Corpus comparables

## Bowker & Pearson, 2002

« *Un corpus comparable est composé d'ensembles de textes, dans des langues différentes, qui ne sont pas des traductions mutuelles.* »

## Déjean & Gaussier, 2002

« *Deux corpus de deux langues $l_1$ et $l_2$ sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue $l_1$, respectivement $l_2$, dont la traduction se trouve dans le corpus de langue $l_2$, respectivement $l_1$.* »

# Approches contextuelles

Rapp (1995) et Fung (1995) introduisent l'alignement à partir de corpus *non-parallèles*. Tous deux s'appuient sur l'idée de caractériser le *contexte* des mots à traduire, plutôt que des informations sur leurs positions.

# Approches contextuelles

Rapp (1995) et Fung (1995) introduisent l'alignement à partir de corpus *non-parallèles*. Tous deux s'appuient sur l'idée de caractériser le *contexte* des mots à traduire, plutôt que des informations sur leurs positions.

# Approches contextuelles

Rapp (1995) et Fung (1995) introduisent l'alignement à partir de corpus *non-parallèles*. Tous deux s'appuient sur l'idée de caractériser le *contexte* des mots à traduire, plutôt que des informations sur leurs positions.

- Fung (1995) s'appuie sur les bigrammes (hétérogénéité à gauche/à droite), Rapp (1995) s'appuie sur les voisins rencontrés dans une fenêtre de taille fixe autour du mot à traduire

# Approches contextuelles

Rapp (1995) et Fung (1995) introduisent l'alignement à partir de corpus *non-parallèles*. Tous deux s'appuient sur l'idée de caractériser le *contexte* des mots à traduire, plutôt que des informations sur leurs positions.

- Fung (1995) s'appuie sur les bigrammes (hétérogénéité à gauche/à droite), Rapp (1995) s'appuie sur les voisins rencontrés dans une fenêtre de taille fixe autour du mot à traduire
- Approche par traduction directe (Fung, 1998)

# Approches contextuelles

Rapp (1995) et Fung (1995) introduisent l'alignement à partir de corpus *non-parallèles*. Tous deux s'appuient sur l'idée de caractériser le *contexte* des mots à traduire, plutôt que des informations sur leurs positions.

- Fung (1995) s'appuie sur les bigrammes (hétérogénéité à gauche/à droite), Rapp (1995) s'appuie sur les voisins rencontrés dans une fenêtre de taille fixe autour du mot à traduire
- Approche par traduction directe (Fung, 1998)
- Approche par similarité interlangue (Déjean & Gaussier, 2002)

# Approches contextuelles

Rapp (1995) et Fung (1995) introduisent l'alignement à partir de corpus *non-parallèles*. Tous deux s'appuient sur l'idée de caractériser le *contexte* des mots à traduire, plutôt que des informations sur leurs positions.

- Fung (1995) s'appuie sur les bigrammes (hétérogénéité à gauche/à droite), Rapp (1995) s'appuie sur les voisins rencontrés dans une fenêtre de taille fixe autour du mot à traduire
- Approche par traduction directe (Fung, 1998)
- Approche par similarité interlangue (Déjean & Gaussier, 2002)

**Firth, 1957**

« On reconnaît un mot à ses fréquentations »

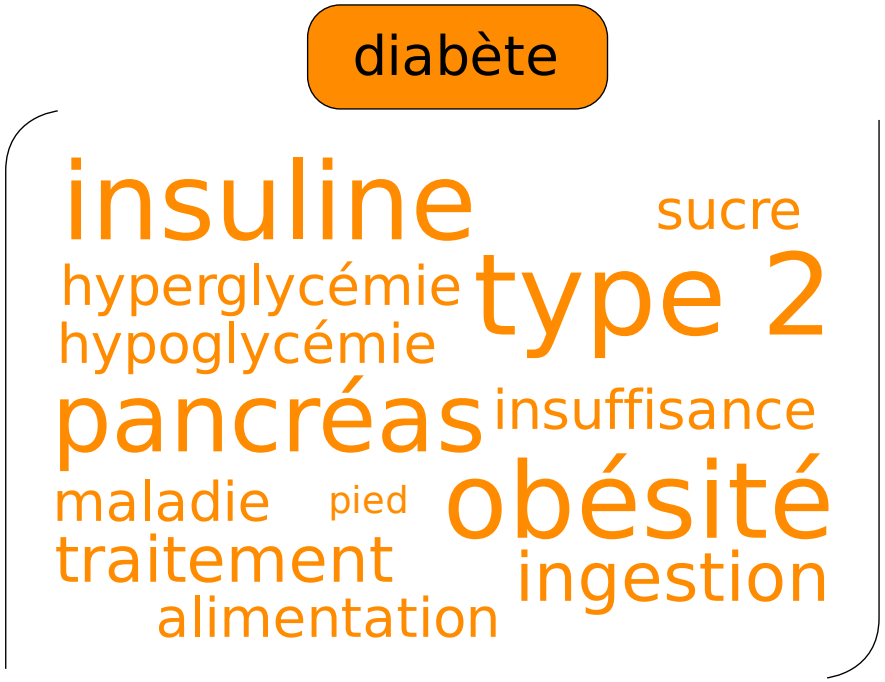
# Approche par traduction directe

1. Construction de *vecteurs de contexte*

eints d'un authentique diabète de type 1 à l'âge adulte p
ante du traitement du diabète traité par insuline. Les dia
rsque l'aggravation du diabète nécessite une escalade th
patients atteints d'un diabète de type 2 nécessitent une
(3,4 mmol/l) lorsque le diabète est associé à 2 facteurs d
majorité atteints d'un diabète de type 2, et leur prise en
r faire le diagnostic de diabète sucré. Il n' est pas recomm
rmale à 3 mois [7]. Le diabète de type 2 s' inscrit habitu
sque de développer le diabète sucré dans environ 50 %
éduisaient le risque de diabète de 58 % lorsqu' elles sont
surveiller l'équilibre du diabète sauf dans les situations a
ait la mortalité liée au diabète de 42 %, la mortalité tout
nt de la découverte du diabète (des complications peuve

# Approche par traduction directe

1. Construction de *vecteurs de contexte*



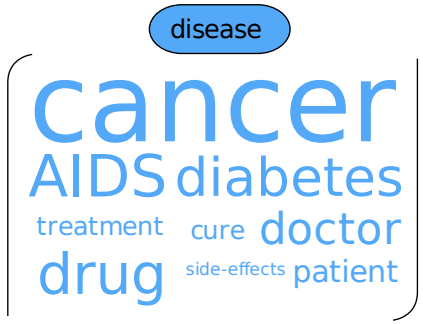
# Approche par traduction directe

1. Construction de *vecteurs de contexte*
2. Traduction des vecteurs sources vers la langue cible

# Approche par traduction directe

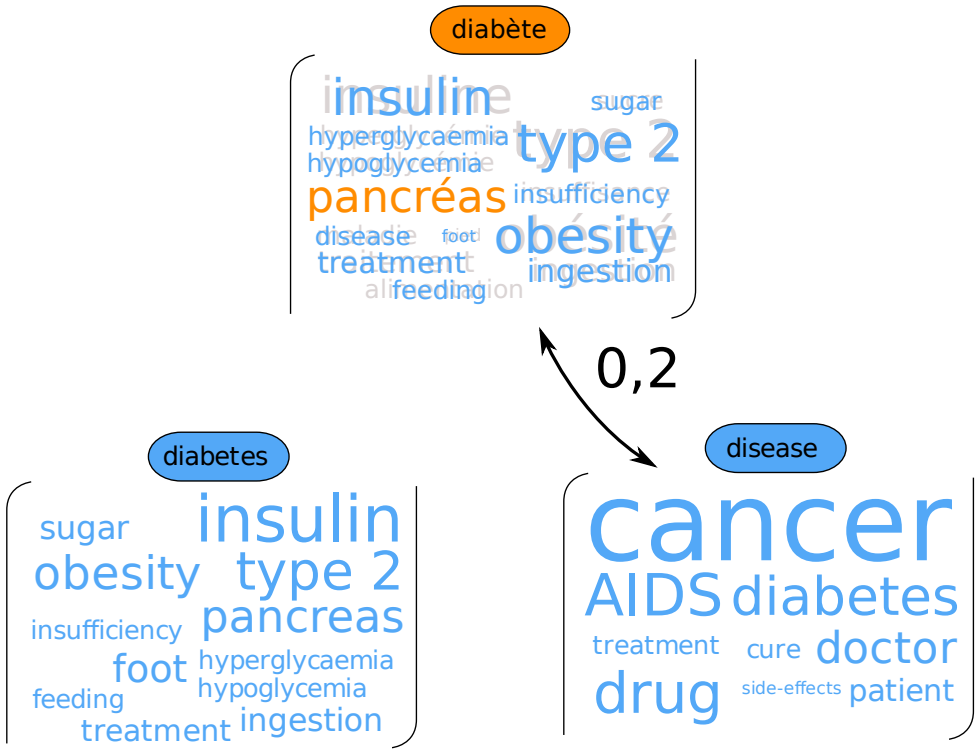
1. Construction de *vecteurs de contexte*
2. Traduction des vecteurs sources vers la langue cible
3. Calcul de la similarité entre vecteurs traduits et cibles





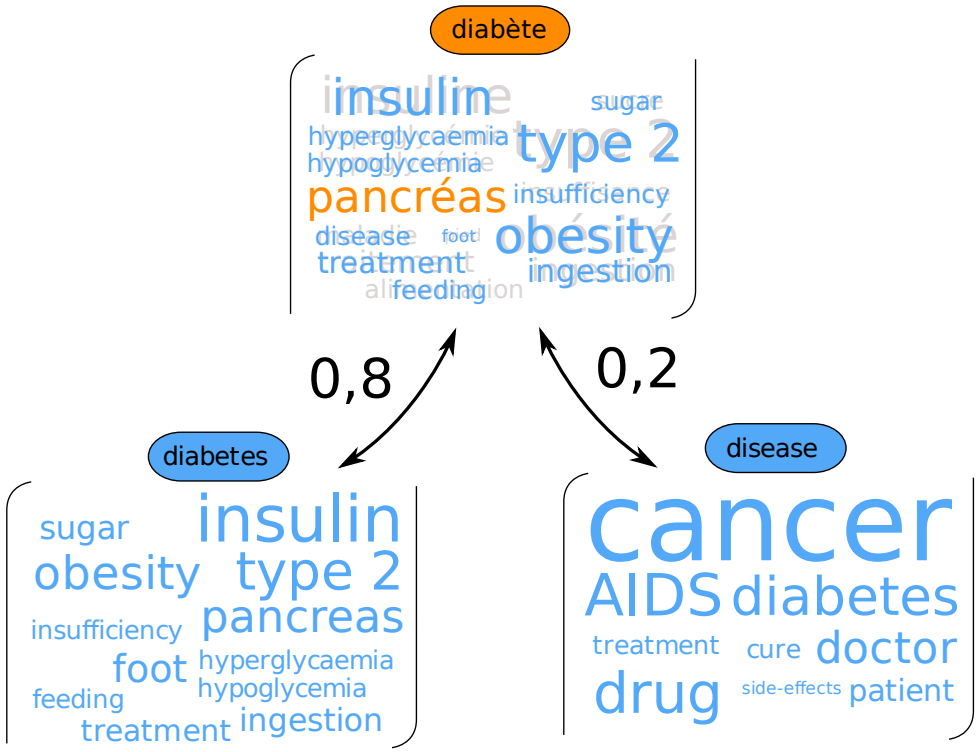
# Approche par traduction directe

1. Construction de *vecteurs de contexte*
2. Traduction des vecteurs sources vers la langue cible
3. Calcul de la similarité entre vecteurs traduits et cibles



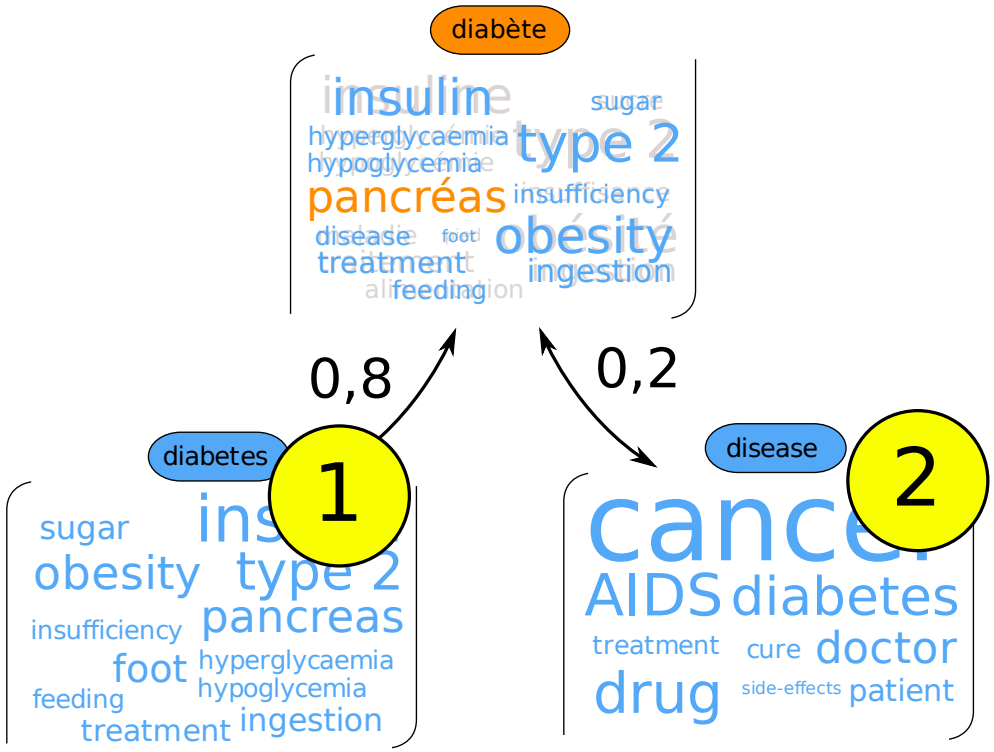
# Approche par traduction directe

1. Construction de *vecteurs de contexte*
2. Traduction des vecteurs sources vers la langue cible
3. Calcul de la similarité entre vecteurs traduits et cibles



# Approche par traduction directe

1. Construction de *vecteurs de contexte*
2. Traduction des vecteurs sources vers la langue cible
3. Calcul de la similarité entre vecteurs traduits et cibles
4. → Liste ordonnée de candidats à la traduction



# Emphase : vecteurs de contexte

Rassemble les cooccurrences *significatives* entre la tête d'un vecteur et ses éléments

**Espace vectoriel** : ensemble des mots collectés dans le corpus (monolingue, source ou cible)

**Coordonnées** : score d'association entre la tête du vecteur et les mots du corpus

# Emphase : vecteurs de contexte

Rassemble les cooccurrences *significatives* entre la tête d'un vecteur et ses éléments

**Espace vectoriel** : ensemble des mots collectés dans le corpus (monolingue, source ou cible)

**Coordonnées** : score d'association entre la tête du vecteur et les mots du corpus

**Association** : indépendance statistique entre deux évènements

*Force de la relation entre deux mots : l'apparition de l'un favorise-t-elle l'apparition de l'autre ? → Sont-ils sémantiquement reliés ?*

# Emphase : vecteurs de contexte

Rassemble les cooccurrences *significatives* entre la tête d'un vecteur et ses éléments

**Espace vectoriel** : ensemble des mots collectés dans le corpus (monolingue, source ou cible)

**Coordonnées** : score d'association entre la tête du vecteur et les mots du corpus

**Association** : indépendance statistique entre deux évènements

*Force de la relation entre deux mots : l'apparition de l'un favorise-t-elle l'apparition de l'autre ? → Sont-ils sémantiquement reliés ?*

Exemple : l'*Information Mutuelle* :

# Emphase : vecteurs de contexte

Rassemble les cooccurrences *significatives* entre la tête d'un vecteur et ses éléments

**Espace vectoriel** : ensemble des mots collectés dans le corpus (monolingue, source ou cible)

**Coordonnées** : score d'association entre la tête du vecteur et les mots du corpus

**Association** : indépendance statistique entre deux évènements

*Force de la relation entre deux mots : l'apparition de l'un favorise-t-elle l'apparition de l'autre ? → Sont-ils sémantiquement reliés ?*

Exemple : l'*Information Mutuelle* :

$IM = log \frac{O}{E}$ ($O$ : observation, $E$ : valeur attendue sous l'hypothèse nulle)

# Emphase : vecteurs de contexte

Rassemble les cooccurrences *significatives* entre la tête d'un vecteur et ses éléments

**Espace vectoriel** : ensemble des mots collectés dans le corpus (monolingue, source ou cible)

**Coordonnées** : score d'association entre la tête du vecteur et les mots du corpus

**Association** : indépendance statistique entre deux évènements

*Force de la relation entre deux mots : l'apparition de l'un favorise-t-elle l'apparition de l'autre ?* $\rightarrow$ *Sont-ils sémantiquement reliés ?*

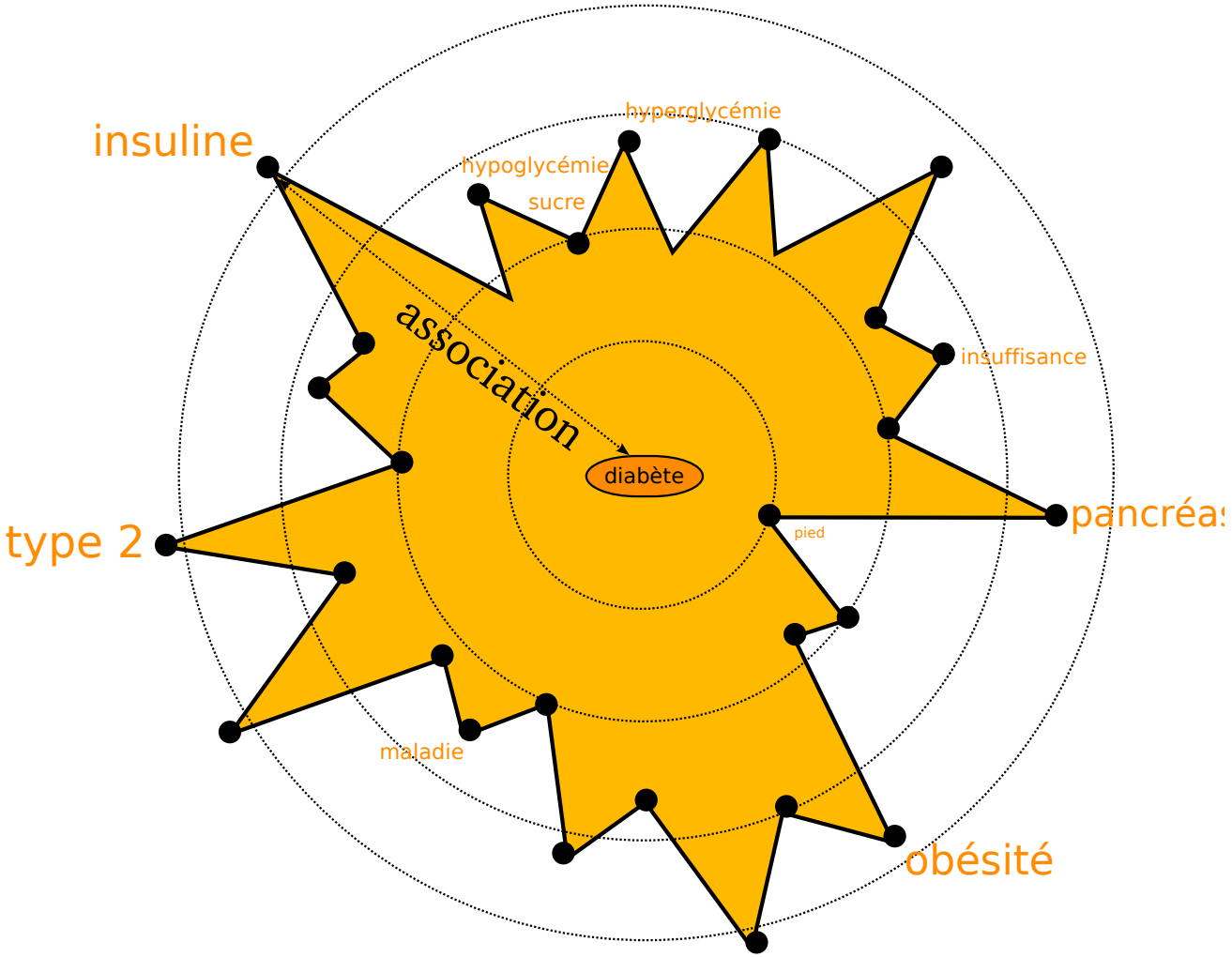
Exemple : l'*Information Mutuelle* :

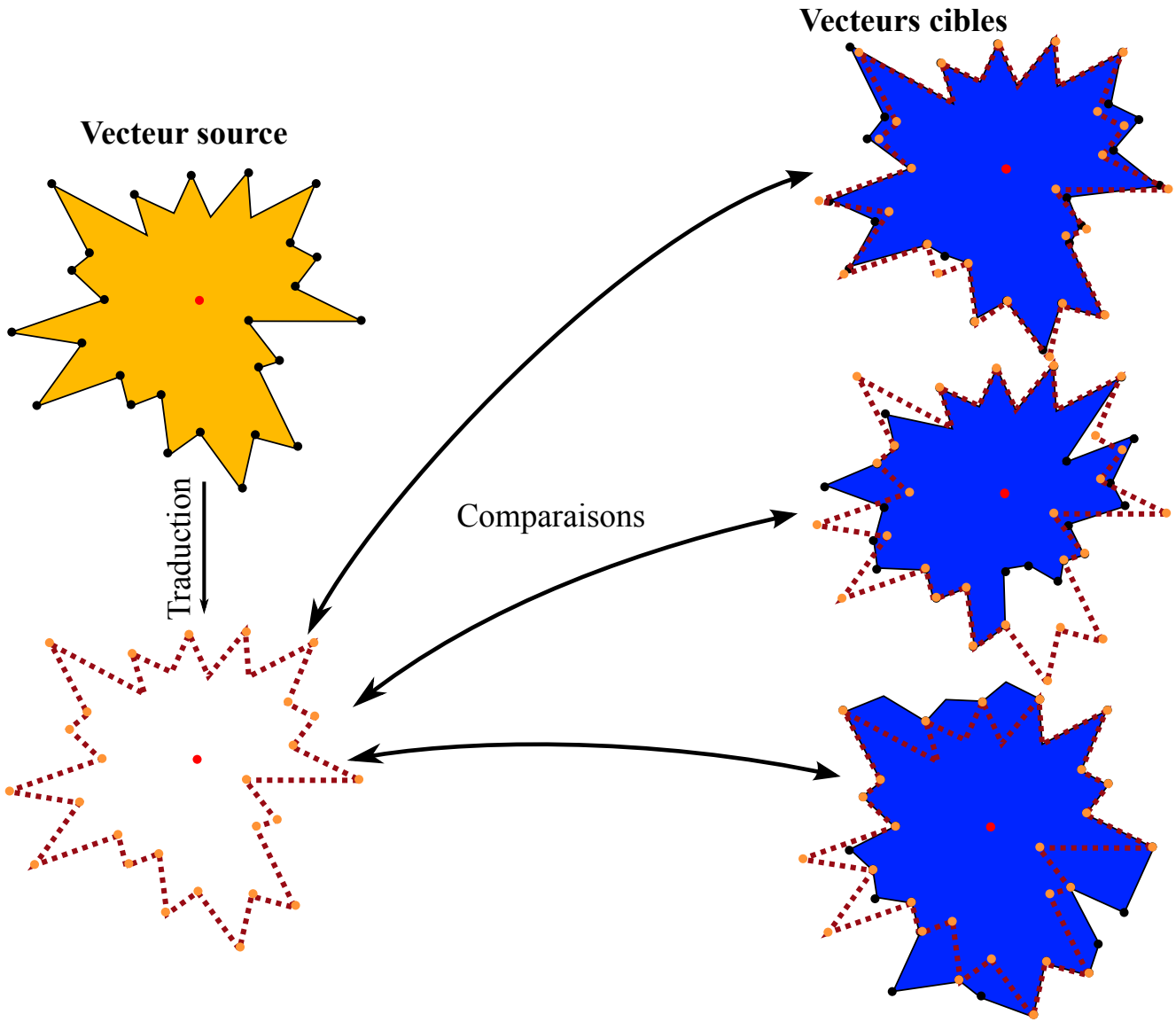
$IM = log \frac{O}{E}$ ($O$ : observation, $E$ : valeur attendue sous l'hypothèse nulle)

$\rightarrow$ **Obtention d'un *Motif d'Association*, pour un mot et ses voisins**

Vecteur de contexte de diabète :

$$\begin{pmatrix} 0 \\ 10{,}04 \\ 0 \\ 0 \\ 3{,}13 \\ 0{,}4 \\ \vdots \\ 0 \\ 0{,}5 \\ 6{,}3 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$




Vecteurs cibles
Vecteur source
Traduction
Comparaisons

# État de l'art

Quelques exemples de travaux connexes :

- Choix des ressources multilingues pour la traduction (Chiao & Zweigenbaum, 2003 ; Déjean & Gaussier, 2002)
- Hypothèse de symétrie distributionnelle (Chiao, 2004)
- Contraintes syntaxiques et lexicales (Sadat *et al.*, 2003)
- Traduction de termes peu fréquents (Pekar *et al.*, 2006)
- Approches géométriques (Gaussier *et al.*, 2004)

# Contexte

- Corpus comparable anglais/français
  - Thème : *cancer du sein*
  - Type de discours *scientifique*
  - 530 000 mots par partie
- Corpus comparable anglais/français/japonais
  - Thème : *diabète et alimentation*
  - Type de discours *scientifique*
  - Environ 250 000 mots par partie

# Problématique

Travail sur des textes spécialisés

# Problématique

Travail sur des textes spécialisés

→ extraction *terminologique* bilingue

# Problématique

Travail sur des textes spécialisés

→ extraction *terminologique* bilingue

Travail sur des corpus de tailles modestes

# Problématique

Travail sur des textes spécialisés

→ extraction *terminologique* bilingue

Travail sur des corpus de tailles modestes

Travail sur des langues distantes

# Paramètres de l'approche directe

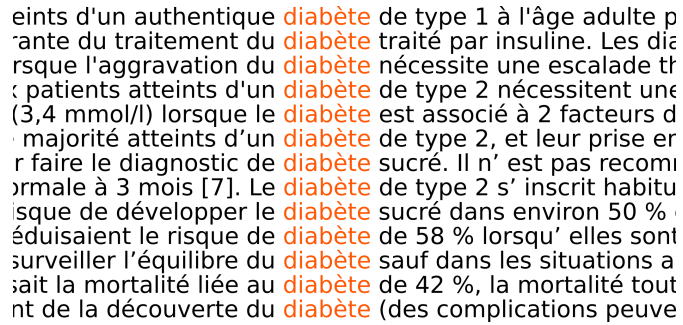
Paramètres principaux :

- Taille de fenêtre contextuelle
- Mesure d'association : *information mutuelle ponctuelle* ou *taux de vraisemblance*
- Mesure de similarité : *cosinus* ou *jaccard pondéré*

# Paramètres de l'approche directe

Paramètres principaux :

- Taille de fenêtre contextuelle
- Mesure d'association : *information mutuelle ponctuelle* ou *taux de vraisemblance*
- Mesure de similarité : *cosinus* ou *jaccard pondéré*

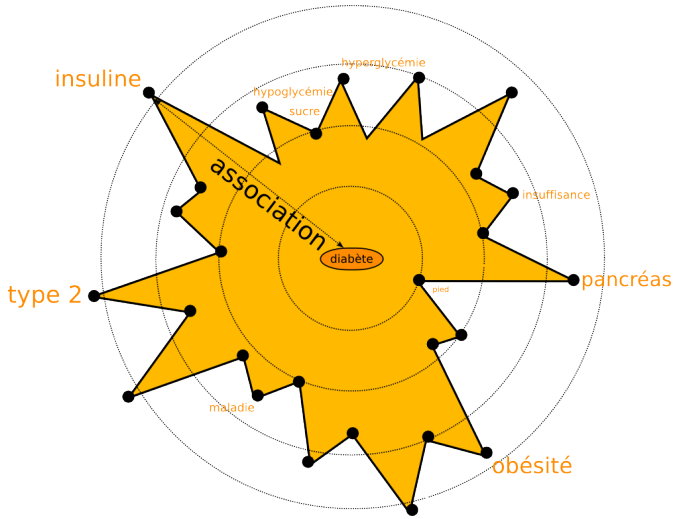



Définit le nombre de voisins à prendre en compte dans le vecteur d'un contexte d'un mot

# Paramètres de l'approche directe

Paramètres principaux :

- Taille de fenêtre contextuelle
- Mesure d'association : *information mutuelle ponctuelle* ou *taux de vraisemblance*
- Mesure de similarité : *cosinus* ou *jaccard pondéré*

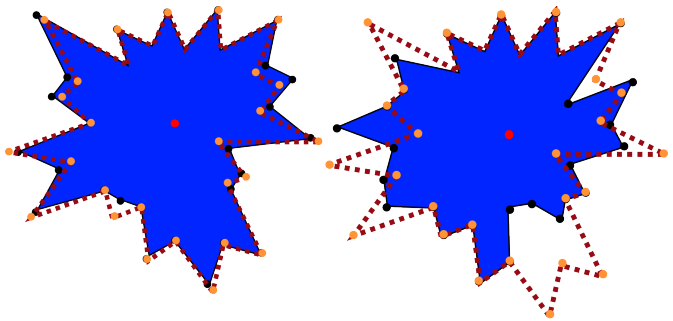



Indique la force de la relation entre deux mots

# Paramètres de l'approche directe

Paramètres principaux :

- Taille de fenêtre contextuelle
- Mesure d'association : *information mutuelle ponctuelle* ou *taux de vraisemblance*
- Mesure de similarité : *cosinus* ou *jaccard pondéré*



Permet de comparer les vecteurs de contexte

# Paramètres de l'approche directe

Paramètres principaux :

- Taille de fenêtre contextuelle
- Mesure d'association : *information mutuelle ponctuelle* ou *taux de vraisemblance*
- Mesure de similarité : *cosinus* ou *jaccard pondéré*
- Paramètres optimaux difficiles à connaître *a priori*

?

# Propositions

Trois propositions :

# Propositions

Trois propositions :

**Sur l'alignement anglais/français → japonais :**

- Exploitation de points d'ancrage pour l'alignement
- Alignement multisources

# Propositions

Trois propositions :

**Sur l'alignement anglais/français → japonais :**

- Exploitation de points d'ancrage pour l'alignement
- Alignement multisources

**Sur l'alignement anglais → français :**

- Exploitation de dépendances syntagmatiques et paradigmatiques en fonction d'indices de fréquence

# Propositions

1. Détection et exploitation de points d'ancrage
2. Fréquences et tailles de fenêtres
3. Alignement multisources

# Points d'ancrage

**Idée : compenser la faible représentativité des données en s'appuyant sur des *éléments de confiance***

# Points d'ancrage

**Idée : compenser la faible représentativité des données en s'appuyant sur des *éléments de confiance*, les points d'ancrage**

# Points d'ancrage

**Idée : compenser la faible représentativité des données en s'appuyant sur des *éléments de confiance*, les points d'ancrage**

## Propriétés

1. Pertinents vis-à-vis des thématiques du corpus
2. Détectables automatiquement
3. Traductions stables

# Points d'ancrage

**Idée : compenser la faible représentativité des données en s'appuyant sur des *éléments de confiance*, les points d'ancrage**

## Propriétés

1. Pertinents vis-à-vis des thématiques du corpus
2. Détectables automatiquement
3. Traductions stables
4. *Couples de traductions*

# Points d'ancrage

**Idée : compenser la faible représentativité des données en s'appuyant sur des *éléments de confiance*, les points d'ancrage**

## Propriétés

1. Pertinents vis-à-vis des thématiques du corpus
2. Détectables automatiquement
3. Traductions stables
4. *Couples de traductions*

S'appuyer sur les points d'ancrage pour :

- $\longrightarrow$ rapprocher les vecteurs traductions
- $\longrightarrow$ éloigner les vecteurs non-traductions

# Points d'ancrage

**Idée : compenser la faible représentativité des données en s'appuyant sur des *éléments de confiance*, les points d'ancrage**

## Propriétés

1. Pertinents vis-à-vis des thématiques du corpus
2. Détectables automatiquement
3. Traductions stables
4. *Couples de traductions*

S'appuyer sur les points d'ancrage pour :

- $\longrightarrow$ rapprocher les vecteurs traductions
- $\longrightarrow$ éloigner les vecteurs non-traductions
- *Rendre les vecteurs de contexte plus discriminants*

# Exploitation des points d'ancrage

**Idée : construire le *motif d'association* en priorité sur les points d'ancrage, puis sur les autres éléments**

# Exploitation des points d'ancrage

**Idée : construire le *motif d'association* en priorité sur les points d'ancrage, puis sur les autres éléments**

→ « Déformation » du motif d'association en faveur des points d'ancrage

# Exploitation des points d'ancrage

**Idée : construire le *motif d'association* en priorité sur les points d'ancrage, puis sur les autres éléments**

$\rightarrow$ « Déformation » du motif d'association en faveur des points d'ancrage

Pour $j$, élément du vecteur $V$ et $PA$ l'ensemble des points d'ancrage pour une langue :

$$assoc\_pondérée_V(j) = \begin{cases} assoc_V(j) + \beta, & \text{si } j \in PA \\ assoc_V(j) - décalage_V, & \text{sinon} \end{cases}$$

# Exploitation des points d'ancrage

**Idée : construire le *motif d'association* en priorité sur les points d'ancrage, puis sur les autres éléments**

$\rightarrow$ « Déformation » du motif d'association en faveur des points d'ancrage

Pour $j$, élément du vecteur $V$ et $PA$ l'ensemble des points d'ancrage pour une langue :

$$assoc\_pondérée_V(j) = \begin{cases} assoc_V(j) + \beta, & \text{si } j \in PA \\ assoc_V(j) - décalage_V, & \text{sinon} \end{cases}$$

Avec $décalage_V = \frac{|V \cap PA|}{|V - PA|} \times \beta$

# Points d'ancrage (2)

Deux types de points d'ancrage identifiés, respectant les propriétés précédentes :

- Translittération : adaptation phonétique d'un mot aux contraintes du japonais

- Composés savants : mots construits sur des racines grecques et latines (Namer, 2005)

# Points d'ancrage (2)

Deux types de points d'ancrage identifiés, respectant les propriétés précédentes :

- Translittération : adaptation phonétique d'un mot aux contraintes du japonais
  インスリン/i-n-su-ri-n (insulin/insuline)

- Composés savants : mots construits sur des racines grecques et latines (Namer, 2005)
  *psychologie*, construit avec le préfixe *psycho-* et le suffixe *-logie*

# Points d'ancrage (2)

Deux types de points d'ancrage identifiés, respectant les propriétés précédentes :

- Translittération : adaptation phonétique d'un mot aux contraintes du japonais
  インスリン/i-n-su-ri-n (insulin/insuline)
  - Facile à identifier automatiquement (syllabaire dédié), nombreuses recherches en alignement automatique

- Composés savants : mots construits sur des racines grecques et latines (Namer, 2005)
  *psychologie*, construit avec le préfixe *psycho-* et le suffixe *-logie*
  - Détectés à l'aide d'une liste d'affixes sur les ressources linguistiques (expression rationnelle)

# Points d'ancrage (2)

Deux types de points d'ancrage identifiés, respectant les propriétés précédentes :

- Translittération : adaptation phonétique d'un mot aux contraintes du japonais

  インスリン/i-n-su-ri-n (insulin/insuline)

  - Facile à identifier automatiquement (syllabaire dédié), nombreuses recherches en alignement automatique
  - Couvre un vocabulaire spécifique, dans le cas des documents *scientifiques* (Ito, 2007)

- Composés savants : mots construits sur des racines grecques et latines (Namer, 2005)

  *psychologie*, construit avec le préfixe *psycho-* et le suffixe *-logie*

  - Détectés à l'aide d'une liste d'affixes sur les ressources linguistiques (expression rationnelle)
  - Caractéristique d'un vocabulaire *scientifique*

# Protocole

Trois expériences sur le corpus *diabète et alimentation*, anglais/français/japonais
Alignements anglais → japonais, français → japonais

1. « *Témoin* »
2. Translittérations
3. Composés savants

# Protocole

Trois expériences sur le corpus *diabète et alimentation*, anglais/français/japonais
Alignements anglais → japonais, français → japonais

1. « *Témoin* »
2. Translittérations – 589 (en/jp) 526 (fr/jp)
3. Composés savants – 604 (en/jp) 819 (fr/jp)

# Protocole

Trois expériences sur le corpus *diabète et alimentation*, anglais/français/japonais
Alignements anglais → japonais, français → japonais

1. « *Témoin* »
2. Translittérations – 589 (en/jp) 526 (fr/jp)
3. Composés savants – 604 (en/jp) 819 (fr/jp)

Paramètres :

# Protocole

Trois expériences sur le corpus *diabète et alimentation*, anglais/français/japonais
Alignements anglais → japonais, français → japonais

1. « *Témoin* »
2. Translittérations – 589 (en/jp) 526 (fr/jp)
3. Composés savants – 604 (en/jp) 819 (fr/jp)

Paramètres :

- Liste de référence de 98 termes
- Taille de fenêtre : 25
- Mesure d'association/Similarité : taux de vraisemblance/cosinus

# Résultats

| | *Témoin* |
|---|---|
| Anglais/Japonais ($Top_1$) | 17,1 % |
| Anglais/Japonais ($Top_{10}$) | 36,3 % |
| Français/Japonais ($Top_1$) | 20,4 % |
| Français/Japonais ($Top_{10}$) | 36,7 % |

Tab.: Résultats de l'alignement anglais-japonais et français-japonais

# Résultats

| | *Témoin* | *Translittérations* |
|---|---|---|
| Anglais/Japonais ($Top_1$) | 17,1 % | 20,2 % [+18,2 %] |
| Anglais/Japonais ($Top_{10}$) | 36,3 % | 39,3 % [+ 8,2 %] |
| Français/Japonais ($Top_1$) | 20,4 % | 20,4 % [+ 0,0 %] |
| Français/Japonais ($Top_{10}$) | 36,7 % | 37,8 % [+ 2,8 %] |

Tab.: Résultats de l'alignement anglais-japonais et français-japonais

# Résultats

| | *Témoin* | *Translittérations* | *Composés Savants* |
|---|---|---|---|
| Anglais/Japonais ($Top_1$) | 17,1 % | 20,2 % [+18,2 %] | 20,2 % [+18,2 %] |
| Anglais/Japonais ($Top_{10}$) | 36,3 % | 39,3 % [+ 8,2 %] | 40,4 % [+11,2 %] |
| Français/Japonais ($Top_1$) | 20,4 % | 20,4 % [+ 0,0 %] | 22,4 % [+10,0 %] |
| Français/Japonais ($Top_{10}$) | 36,7 % | 37,8 % [+ 2,8 %] | 38,8 % [+ 5,6 %] |

Tab.: Résultats de l'alignement anglais-japonais et français-japonais

# Analyse

Effet des points d'ancrage sur les résultats de l'alignement :

- $\rightarrow$ léger reclassement des candidats bien classés ($Top < 15$)
- $\rightarrow$ large reclassement des candidats mal classés ($Top > 50$)

# Analyse

Effet des points d'ancrage sur les résultats de l'alignement :

- $\rightarrow$ léger reclassement des candidats bien classés ($Top < 15$)
- $\rightarrow$ large reclassement des candidats mal classés ($Top > 50$)

**Amélioration globale des résultats**

# Analyse

Effet des points d'ancrage sur les résultats de l'alignement :

- $\rightarrow$ léger reclassement des candidats bien classés ($Top < 15$)
- $\rightarrow$ large reclassement des candidats mal classés ($Top > 50$)

**Amélioration globale des résultats $\rightarrow$ même si améliorations $Top_1$ et $Top_{10}$ faibles**

# Analyse

Effet des points d'ancrage sur les résultats de l'alignement :

- $\rightarrow$ léger reclassement des candidats bien classés ($Top < 15$)
- $\rightarrow$ large reclassement des candidats mal classés ($Top > 50$)

**Amélioration globale des résultats $\rightarrow$ même si améliorations $Top_1$ et $Top_{10}$ faibles**

Possibilité d'extension à d'autres langues/d'autres types de vocabulaire

# Analyse

Effet des points d'ancrage sur les résultats de l'alignement :

- $\rightarrow$ léger reclassement des candidats bien classés ($Top < 15$)
- $\rightarrow$ large reclassement des candidats mal classés ($Top > 50$)

**Amélioration globale des résultats $\rightarrow$ même si améliorations $Top_1$ et $Top_{10}$ faibles**

Possibilité d'extension à d'autres langues/d'autres types de vocabulaire

- Phénomène de translittérations présent en Arabe, en Chinois...
- Phénomène de *cognats* entre langues (ex : gouvernement/government)

# Propositions

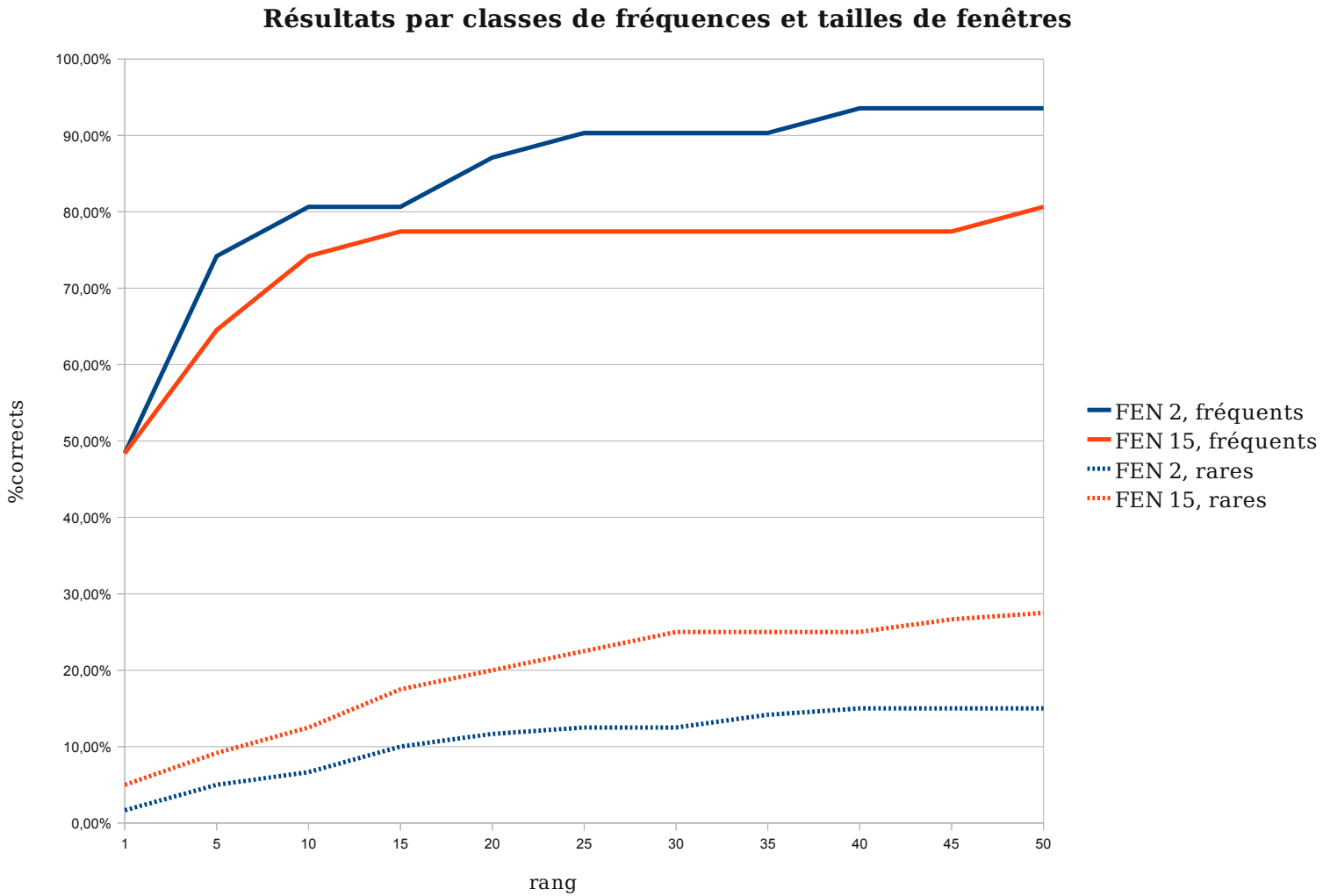
1. Détection et exploitation de points d'ancrage
2. Fréquences et tailles de fenêtres
3. Alignement multisources

# Observations

Observations :

1. Les mots très fréquents sont mieux caractérisés par de petites tailles de fenêtre contextuelle (<3)
2. Les mots plus rares sont mieux caractérisés par de grandes tailles de fenêtre (>15)

# Observations



# Hypothèse

Les mots fréquents peuvent être caractérisés par des dépendances *syntagmatiques* (Grefenstette, 1996)

- Ex : relations sujet-verbe, syntagme nominaux (*onde électromagnétique*)
- → Utilisation d'une information abondante mais très pertinente.

# Hypothèse

Les mots fréquents peuvent être caractérisés par des dépendances *syntagmatiques* (Grefenstette, 1996)

- Ex : relations sujet-verbe, syntagme nominaux (*onde électromagnétique*)
- → Utilisation d'une information abondante mais très pertinente.

Les mots plus rares ne peuvent supporter cette disette d'information (Zweigenbaum & Habert, 2006)

- Nécessité d'absorber un maximum d'information
- Prise en compte des dépendances *paradigmatiques*
- Ex : *diabète . . . pancréas*

# Hypothèse

Les mots fréquents peuvent être caractérisés par des dépendances *syntagmatiques* (Grefenstette, 1996)

- Ex : relations sujet-verbe, syntagme nominaux (*onde électromagnétique*)
- → Utilisation d'une information abondante mais très pertinente.

Les mots plus rares ne peuvent supporter cette disette d'information (Zweigenbaum & Habert, 2006)

- Nécessité d'absorber un maximum d'information
- Prise en compte des dépendances *paradigmatiques*
- Ex : *diabète . . . pancréas*
- → Perte de pertinence au profit de la quantité.

# Proposition

- Prendre en compte la fréquence d'un mot pour savoir comment le caractériser
- Fréquences élevées : petite fenêtre
- Fréquences faibles : grande fenêtre
- Fréquences intermédiaires : fenêtre intermédiaire

# Protocole

Deux expériences :

1. Fenêtre taille fixe (Témoin)
2. Fenêtre variable en fonction de la fréquence

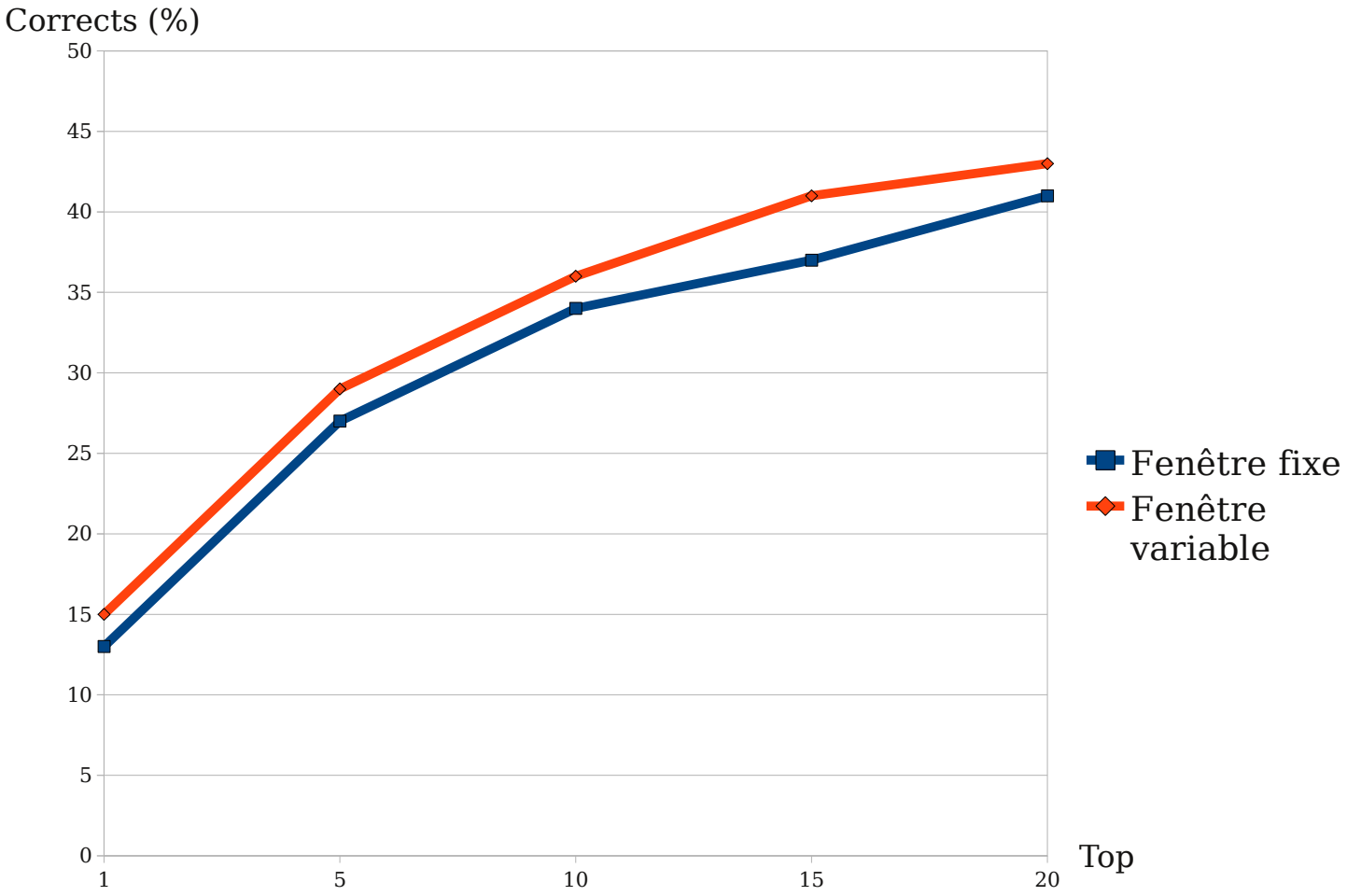
Paramètres :

- Liste de 648 traductions français/anglais
- Mesures d'association/similarité : taux de vraisemblance/jaccard pondéré
- fenêtre taille 3 pour expérience témoin

Table de correspondances :

| Fréquence | Fenêtre |
|---|---|
| [15–25] | 20 |
| ]25–30] | 15 |
| ]30–100] | 10 |
| ]100–200] | 5 |
| ]200–500] | 3 |
| ]500–+∞[ | 2 |

# Résultat


Corrects (%)
50
45
40
35
30
25
20
15
10
5
0
1
5
10
15
20
Top
Fenêtre fixe
Fenêtre variable

# Discussion

**Intérêt de cette méthode :**

Choix *a priori* d'un paramètre...

# Discussion

**Intérêt de cette méthode :**

Choix *a priori* d'un paramètre...

... en fonction d'un indice absolu et facile à obtenir : la fréquence

# Discussion

**Intérêt de cette méthode :**

Choix *a priori* d'un paramètre...

... en fonction d'un indice absolu et facile à obtenir : la fréquence

Mais inefficace dans le cas du japonais

# Discussion

**Intérêt de cette méthode :**

Choix *a priori* d'un paramètre...

... en fonction d'un indice absolu et facile à obtenir : la fréquence

Mais inefficace dans le cas du japonais

→ Meilleurs résultats toujours obtenus avec une grande taille de fenêtre

# Discussion

**Intérêt de cette méthode :**

Choix *a priori* d'un paramètre...

... en fonction d'un indice absolu et facile à obtenir : la fréquence

Mais inefficace dans le cas du japonais

→ Meilleurs résultats toujours obtenus avec une grande taille de fenêtre

Spécificités de la langue japonaise ?

# Propositions

1. Détection et exploitation de points d'ancrage
2. Fréquences et tailles de fenêtres
3. Alignement multisources

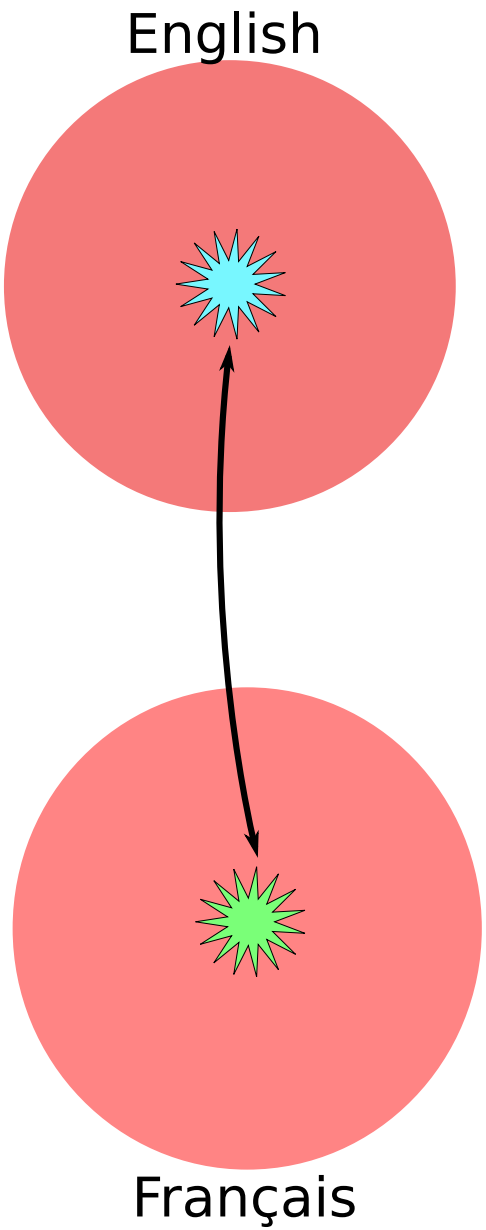
# Contexte :

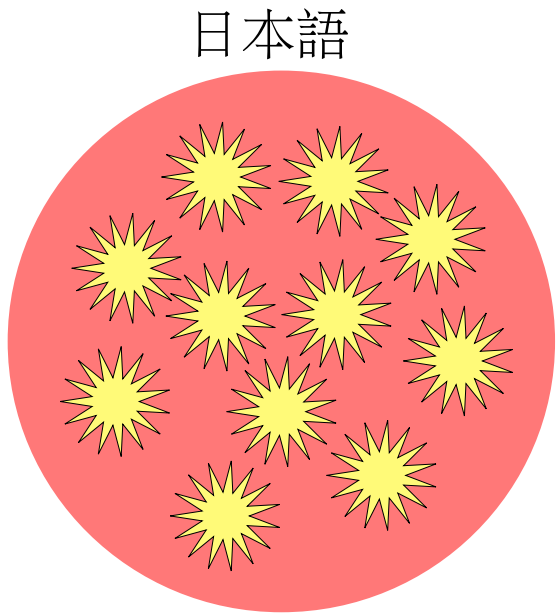
Corpus comparable français-anglais-japonais → 3 langues disponibles

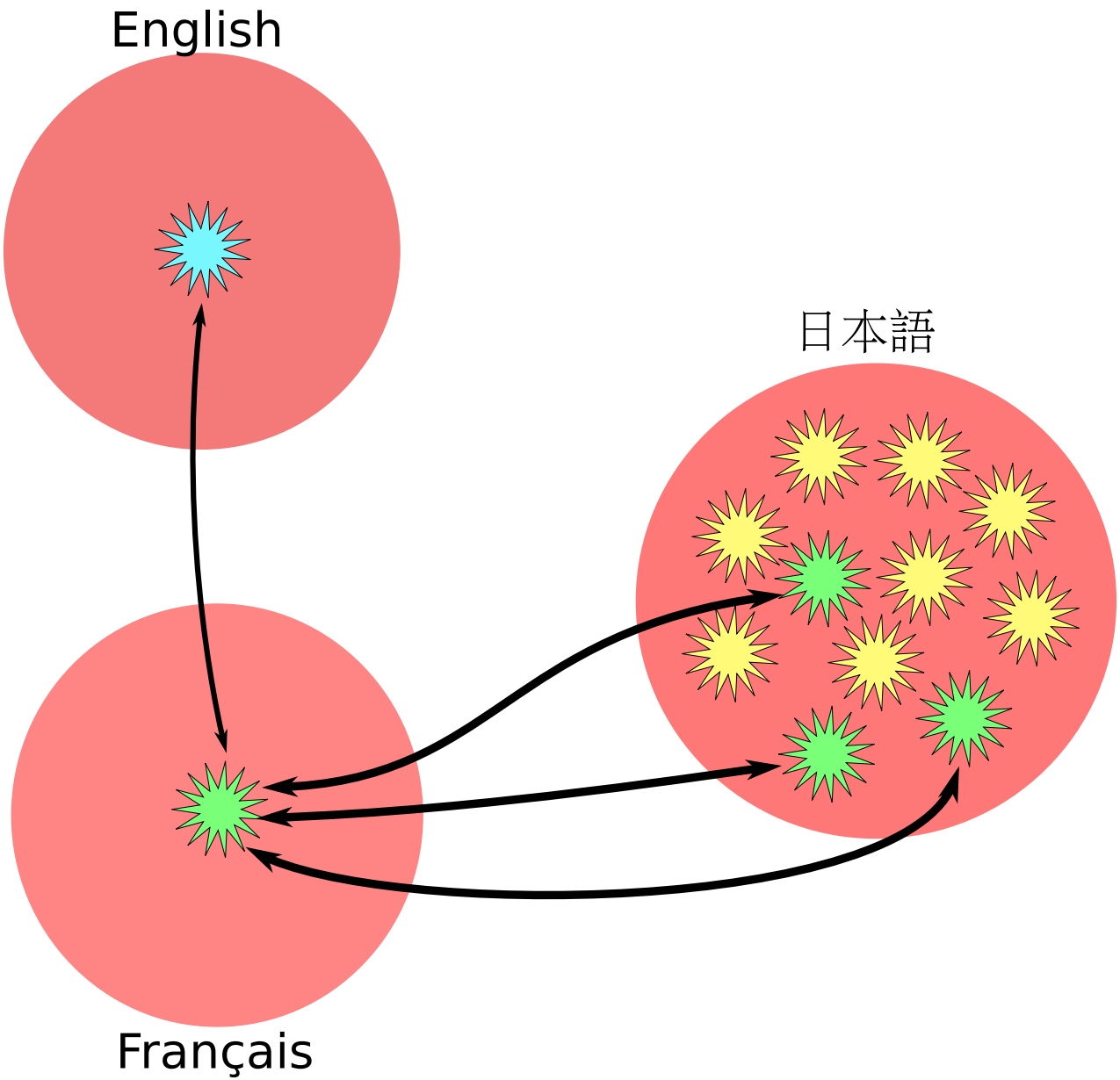
Alignement français-anglais correct + ressources lexicales plus complètes

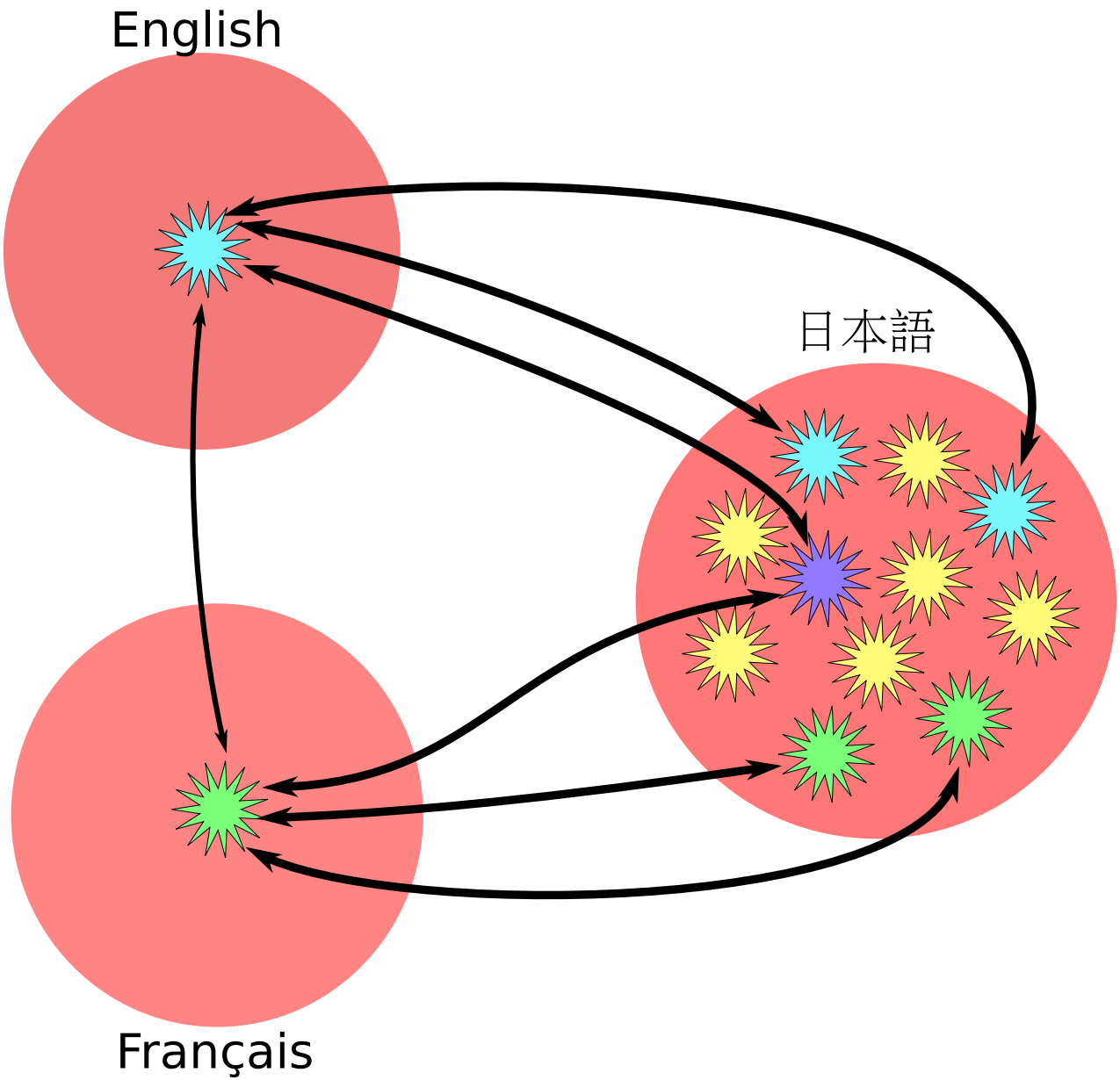
Alignements français-japonais ; anglais-japonais délicat

Idée : exploitation des connaissances en français-anglais pour aligner vers le japonais


English
Français


日本語


English
日本語
Français


English
日本語
Français

# Proposition

Utilisation de la moyenne harmonique pour reclasser les candidats à la traduction

# Proposition

Utilisation de la moyenne harmonique pour reclasser les candidats à la traduction

- $r_{candidat} = MH(r_{en,jp}, r_{fr,jp}) = \frac{2 r_{en,jp} r_{fr,jp}}{r_{en,jp} + r_{fr,jp}}$
- Note : les candidats qui n'apparaissent pas dans les **deux** alignements sont éliminés

# Proposition

Utilisation de la moyenne harmonique pour reclasser les candidats à la traduction

- $r_{candidat} = MH(r_{en,jp}, r_{fr,jp}) = \frac{2 r_{en,jp} r_{fr,jp}}{r_{en,jp} + r_{fr,jp}}$
- Note : les candidats qui n'apparaissent pas dans les **deux** alignements sont éliminés

Ex : recherche de la traduction d'insuline/insulin

# Proposition

Utilisation de la moyenne harmonique pour reclasser les candidats à la traduction

- $r_{candidat} = MH(r_{en,jp}, r_{fr,jp}) = \frac{2 r_{en,jp} r_{fr,jp}}{r_{en,jp} + r_{fr,jp}}$
- Note : les candidats qui n'apparaissent pas dans les **deux** alignements sont éliminés

Ex : recherche de la traduction d'insuline/insulin

| *Rang :* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *insulin* | インスリン | 自律 | 膵臓 | 改善 | 障害 | シンケイ | 代償 | 促進 |
| *insuline* | 促進 | 代償 | インスリン | 分泌 | 注射 | 改善 | 膵臓 | 末梢 |

# Proposition

Utilisation de la moyenne harmonique pour reclasser les candidats à la traduction

- $r_{candidat} = MH(r_{en,jp}, r_{fr,jp}) = \frac{2 r_{en,jp} r_{fr,jp}}{r_{en,jp} + r_{fr,jp}}$
- Note : les candidats qui n'apparaissent pas dans les **deux** alignements sont éliminés

Ex : recherche de la traduction d'insuline/insulin

| *Rang :* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *insulin* | インスリン | 自律 | 膵臓 | 改善 | 障害 | シンケイ | 代償 | 促進 |
| *insuline* | 促進 | 代償 | インスリン | 分泌 | 注射 | 改善 | 膵臓 | 末梢 |

Combinaison des résultats :

`{insuline,insulin}` → インスリン ; 促進 ; 代償 ; 膵臓 ; 改善

# Expérience

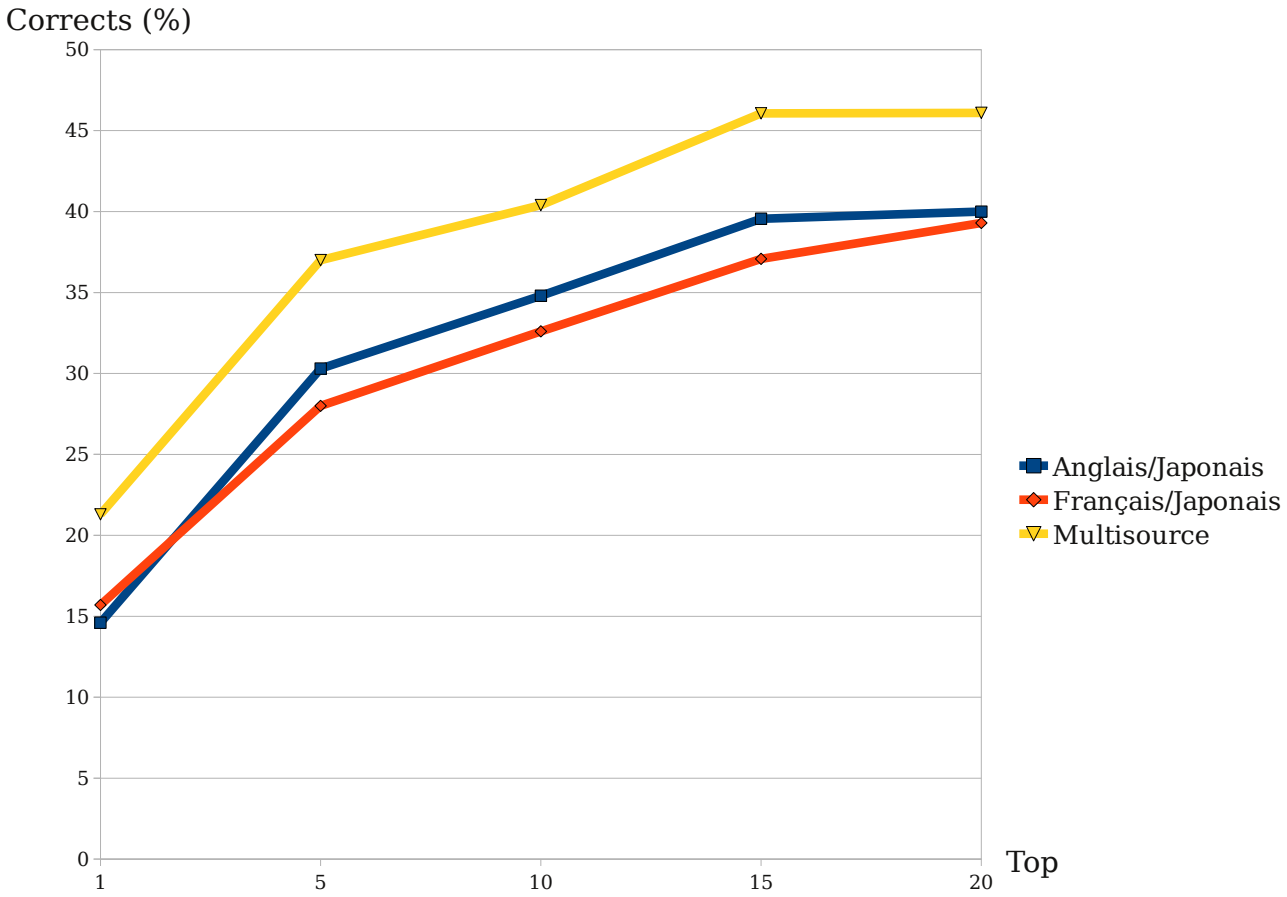
Trois expériences :

1. Témoin français/japonais
2. Témoin anglais/japonais
3. Alignement multisources

Paramètres :

- Intersection de listes de termes français/japonais et anglais/japonais
- Liste de 89 termes en traduction, français/anglais/japonais
- Fenêtre taille 15
- Mesure d'association/similarité : taux de vraisemblance/cosinus

# Résultats


Corrects (%)
50
45
40
35
30
25
20
15
10
5
0
1
5
10
15
20
Top
Anglais/Japonais
Français/Japonais
Multisource

# Discussion

**Intérêt** : fort gain en précision

**Conséquence principale** : facilite l'accès à l'information :

# Discussion

**Intérêt** : fort gain en précision

**Conséquence principale** : facilite l'accès à l'information :

- Difficulté à construire de *gros* corpus fortement comparables

# Discussion

**Intérêt** : fort gain en précision

**Conséquence principale** : facilite l'accès à l'information :

- Difficulté à construire de *gros* corpus fortement comparables
- Facilité à construire des corpus comparables modestes dans de nombreuses langues, sur de nombreux sujets

# Discussion

**Intérêt** : fort gain en précision

**Conséquence principale** : facilite l'accès à l'information :

- Difficulté à construire de *gros* corpus fortement comparables
- Facilité à construire des corpus comparables modestes dans de nombreuses langues, sur de nombreux sujets
- Bonne connaissance de certaines langues

# Discussion

**Intérêt** : fort gain en précision
**Conséquence principale** : facilite l'accès à l'information :

- Difficulté à construire de *gros* corpus fortement comparables
- Facilité à construire des corpus comparables modestes dans de nombreuses langues, sur de nombreux sujets
- Bonne connaissance de certaines langues

Exploitation de langues connues pour des alignements plus délicats

# Discussion

**Intérêt** : fort gain en précision
**Conséquence principale** : facilite l'accès à l'information :

- Difficulté à construire de *gros* corpus fortement comparables
- Facilité à construire des corpus comparables modestes dans de nombreuses langues, sur de nombreux sujets
- Bonne connaissance de certaines langues

Exploitation de langues connues pour des alignements plus délicats
Ouverture : *gros c'est beau* vs. *contraint c'est bien* (Morin *et al.* 2009) → *qualité dans la variété* !

# Comparabilité des corpus comparables

**Comparabilité** : intuitivement, qualité d'un corpus comparable pour une tâche donnée
→ Notion d'homogénéité multilingue (Kilgarriff, 2001)
Dans la littérature, comparabilité évaluée sur :

- Critères qualitatifs/externes : choix des documents composant le corpus (thème, type de discours, date de publication, origine linguistique. . . – Bowker & Pearson, 2002)
- Critères quantitatifs/internes : quantité de vocabulaire commun entre les différentes parties du corpus (Déjean & Gaussier, 2002 ; Saralegi & Alegrìa, 2007)

# Comparabilité

Cas de l'extraction lexicale bilingue

- Critère quantitatif prépondérant → vocabulaire commun
- Généralement, le respect des critères qualitatifs tend à maximiser les critères quantitatifs

**Comparabilité** : intuitivement, l'idée que les parties du corpus se ressemblent, ont des caractéristiques similaires

- Corpus *équilibrés*, pour favoriser une même représentativité du vocabulaire commun

# Comparabilité

Cas de l'extraction lexicale bilingue

- Critère quantitatif prépondérant → vocabulaire commun
- Généralement, le respect des critères qualitatifs tend à maximiser les critères quantitatifs

**Comparabilité** : intuitivement, l'idée que les parties du corpus se ressemblent, ont des caractéristiques similaires

- Corpus *équilibrés*, pour favoriser une même représentativité du vocabulaire commun
  - (Morin, 2009) montre qu'il est possible d'exploiter des corpus massivement déséquilibrés

# Comparabilité

Cas de l'extraction lexicale bilingue

- Critère quantitatif prépondérant → vocabulaire commun
- Généralement, le respect des critères qualitatifs tend à maximiser les critères quantitatifs

**Comparabilité** : intuitivement, l'idée que les parties du corpus se ressemblent, ont des caractéristiques similaires

- Corpus *équilibrés*, pour favoriser une même représentativité du vocabulaire commun
  - (Morin, 2009) montre qu'il est possible d'exploiter des corpus massivement déséquilibrés *à condition de choisir soigneusement les paramètres de l'alignement*

# Comparabilité

Cas de l'extraction lexicale bilingue

- Critère quantitatif prépondérant → vocabulaire commun
- Généralement, le respect des critères qualitatifs tend à maximiser les critères quantitatifs

**Comparabilité** : intuitivement, l'idée que les parties du corpus se ressemblent, ont des caractéristiques similaires

- Corpus *équilibrés*, pour favoriser une même représentativité du vocabulaire commun
  - (Morin, 2009) montre qu'il est possible d'exploiter des corpus massivement déséquilibrés *à condition de choisir soigneusement les paramètres de l'alignement*
- Corpus *propres*, *homogènes*, pour limiter le bruit
  - Vraiment ?

# Corpus bruité

**Hypothèse**

Vues les méthodes utilisées pour l'extraction lexicale bilingue à partir de vecteurs de contexte, le bruit ne devrait pas avoir d'incidence sur les résultats de l'extraction, à condition que ce bruit n'interfère pas avec le vocabulaire à aligner.

Expérience : alignement français-anglais (corpus *cancer du sein*)

- partie française complétée par une partie du corpus issu du journal *Le Monde*
- partie anglaise complétée par un extrait du corpus *EUROPARL*

**Le volume de *bruit* est supérieur au volume initial du corpus**

# Résultats

Observations :

1. Différence de qualité de l'alignement peu significative
2. **L'évolution du rang des traductions trouvées entre l'expérience témoin (corpus équilibré) et l'expérience bruitée est moins significative que celle obtenue en changeant simplement les mesures d'association utilisées pour l'expérience témoin**

# Résultats

Observations :

1. Différence de qualité de l'alignement peu significative
2. **L'évolution du rang des traductions trouvées entre l'expérience témoin (corpus équilibré) et l'expérience bruitée est moins significative que celle obtenue en changeant simplement les mesures d'association utilisées pour l'expérience témoin**

## Hypothèse confirmée

Le bruit dans les corpus n'influence pas significativement l'alignement

# Résultats

Observations :

1. Différence de qualité de l'alignement peu significative
2. **L'évolution du rang des traductions trouvées entre l'expérience témoin (corpus équilibré) et l'expérience bruitée est moins significative que celle obtenue en changeant simplement les mesures d'association utilisées pour l'expérience témoin**

## Hypothèse confirmée

Le bruit dans les corpus n'influence pas significativement l'alignement *à condition que le bruit n'interfère pas avec le lexique à aligner*

# Notion d'incomparabilité

Deux contraintes disparaissent :

- la contrainte d'équilibre des corpus
- la contrainte d'*homogénéité des corpus*

# Notion d'incomparabilité

Deux contraintes disparaissent :

- la contrainte d'équilibre des corpus
- la contrainte d'*homogénéité des corpus* à condition que le bruit ne couvre pas le vocabulaire à extraire

# Notion d'incomparabilité

Deux contraintes disparaissent :

- la contrainte d'équilibre des corpus
- la contrainte d'*homogénéité des corpus* à condition que le bruit ne couvre pas le vocabulaire à extraire

- Exemple : *history*
  - Sens 1 : *antécédent*
  - Sens 2 : *historique, histoire*

# Notion d'incomparabilité

Deux contraintes disparaissent :

- la contrainte d'équilibre des corpus
- la contrainte d'*homogénéité des corpus* à condition que le bruit ne couvre pas le vocabulaire à extraire

- Exemple : *history*
  - Sens 1 : *antécédent*
  - Sens 2 : *historique, histoire*

**Définition : incomparabilité**

Élément lexical utilisé avec des sens différents au sein d'un même corpus

# Comparabilité

- Dépistage et traitement des incomparabilités
- Exemple : distinguer les différents usages de *drug*
  - Drogue, contexte : *addiction, désintoxication, répression. . .*
  - Médicament, contexte : *traitement, prescription, posologie. . .*
- Construire un vecteur de contexte par usage

Connexion avec l'*acquisition sémantique* (Grefenstette, 1994)

La comparabilité repose sur :

+ La quantité de vocabulaire commun
− La quantité d'*incomparabilité*

# Conséquences

- Moins de contraintes sur la construction des corpus
- Meilleures approches pour l'exploitation des corpus dans le cas de l'extraction lexicale bilingue
  - cas des corpus homogènes et équilibrés : approches classiques
  - cas des corpus hétérogènes et/ou déséquilibrés : approches classiques, avec choix pertinent des mesures d'association
  - cas des corpus *incomparables* : partitionnement sémantique *puis* approches classiques

# Conclusion

Enjeux : **Amélioration de la caractérisation terminologique multilingue et de l'alignement de lexique bilingue à partir de corpus comparables**

- Trois améliorations de l'approche directe
  - Utilisation de points d'ancrage
  - Corrélation fréquence d'un mot/taille de fenêtre optimale
  - Alignement multisource
- Discussion sur la notion de comparabilité : cadre, évaluation et conséquences
  - Conséquences sur la façon d'exploiter les corpus
  - Conséquences sur la façon de construire les corpus
- Choix de certains paramètres *a priori* pour l'approche directe
  - Taille de fenêtre (en fonction de la fréquence)
  - Mesure d'association (en fonction de la *comparabilité* du corpus)

# Perspectives

Possibilités d'améliorations de l'extraction de lexique bilingue à partir de corpus comparables spécialisés :

- par combinaison de méthodes, peu exploitées ici
- par une meilleure caractérisation *a priori* des corpus comparables

Exploitation de corpus comparables *différents* :

- corpus variés, pour couvrir plus précisément des domaines plus larges
- problème des incomparabilités : approches plus complexes et plus subtiles

Domaine d'étude en pleine expansion

- atelier sur les corpus comparables (Fung & Zweigenbaum, 2008)
- construction de corpus comparables (Goeuriot, 2008 ; Sharoff, 2006)

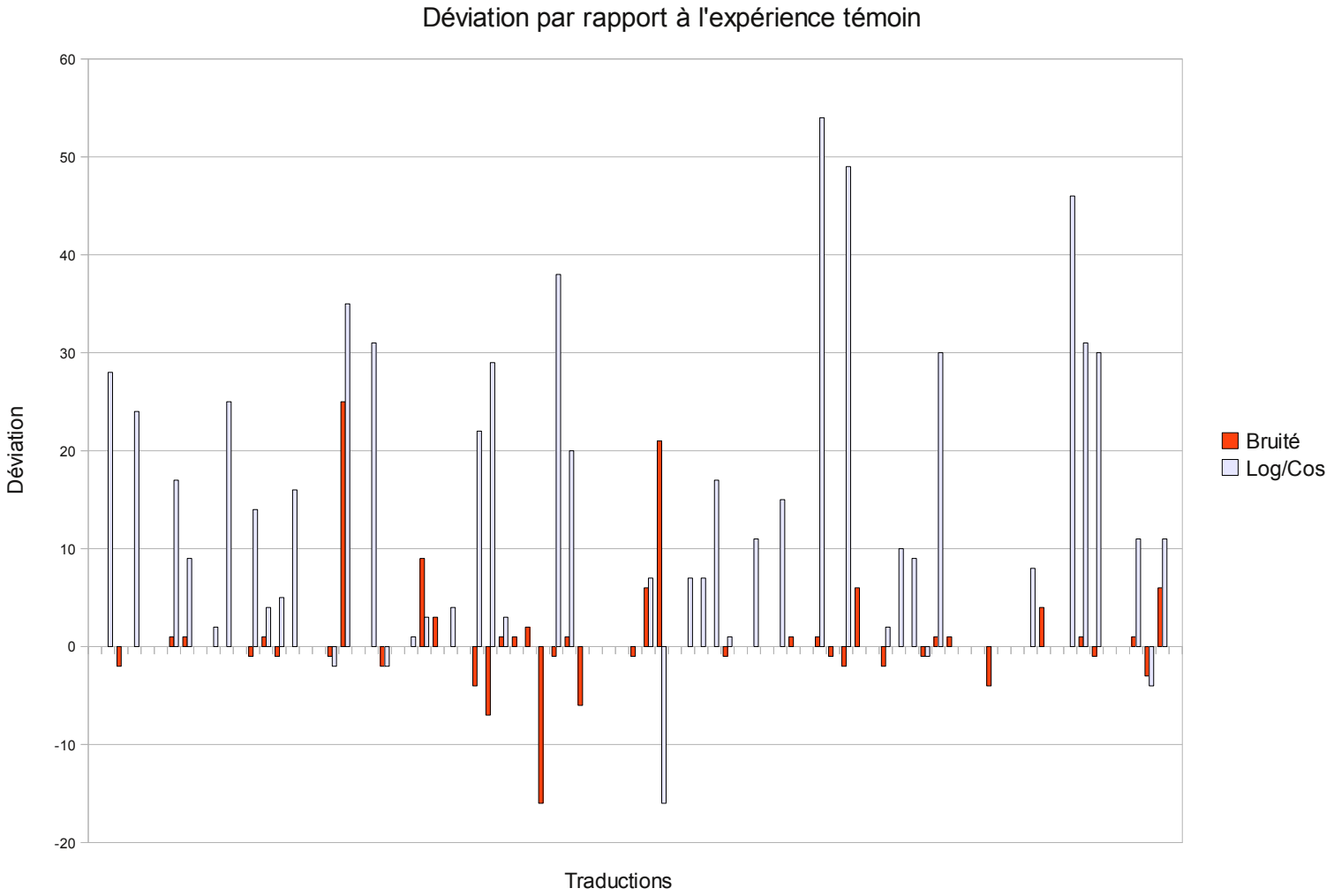
# Merci !

# Merci !

# Merci !

# Merci !

# Merci !


Déviation par rapport à l'expérience témoin
Déviation
60
50
40
30
20
10
0
-10
-20
Traductions
Bruité
Log/Cos