Influence des points d'ancrage pour l'extraction de lexique bilingue à partir de corpus comparables spécialisés

Emmanuel Prochasson
Emmanuel Morin

Laboratoire d'Informatique de Nantes Atlantique

TALN 2009

Introduction

- Objectif : extraire un vocabulaire commun et l'aligner automatiquement pour constituer un lexique bilingue

Introduction

- Objectif : extraire un vocabulaire commun et l'aligner automatiquement pour constituer un lexique bilingue
- à partir de documents (multilingues) n'étant pas en relation de traduction

Introduction

- Objectif : extraire un vocabulaire commun et l'aligner automatiquement pour constituer un lexique bilingue
- à partir de documents (multilingues) n'étant pas en relation de traduction
- $\longrightarrow$ corpus *comparables*

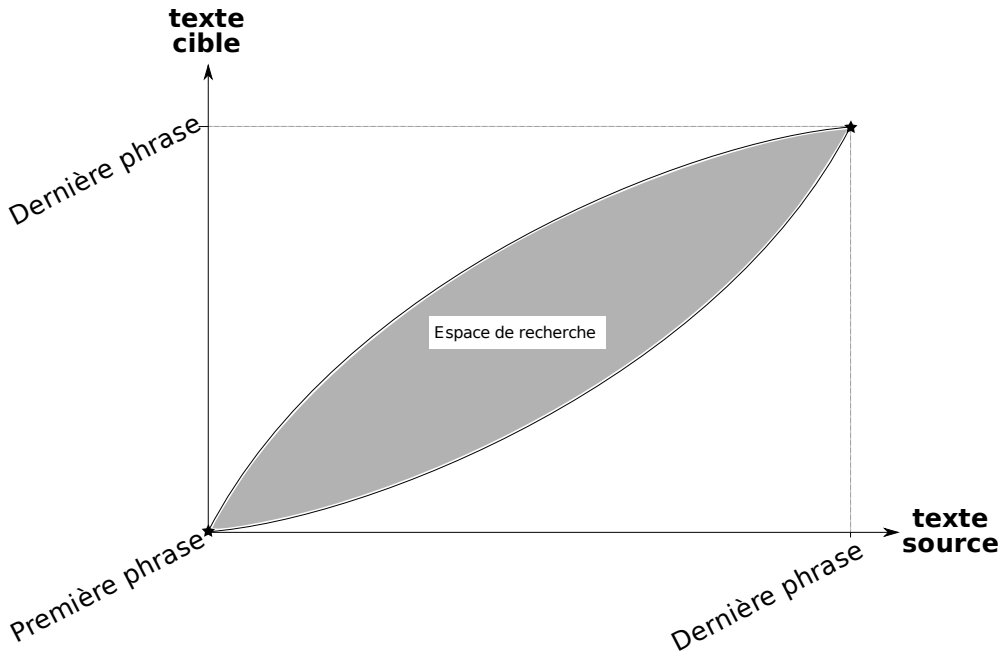
Introduction

- Objectif : extraire un vocabulaire commun et l'aligner automatiquement pour constituer un lexique bilingue
- à partir de documents (multilingues) n'étant pas en relation de traduction
- $\longrightarrow$ corpus *comparables*
- « *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* » (Déjean & Gaussier, 2002)

Plan

Corpus parallèles

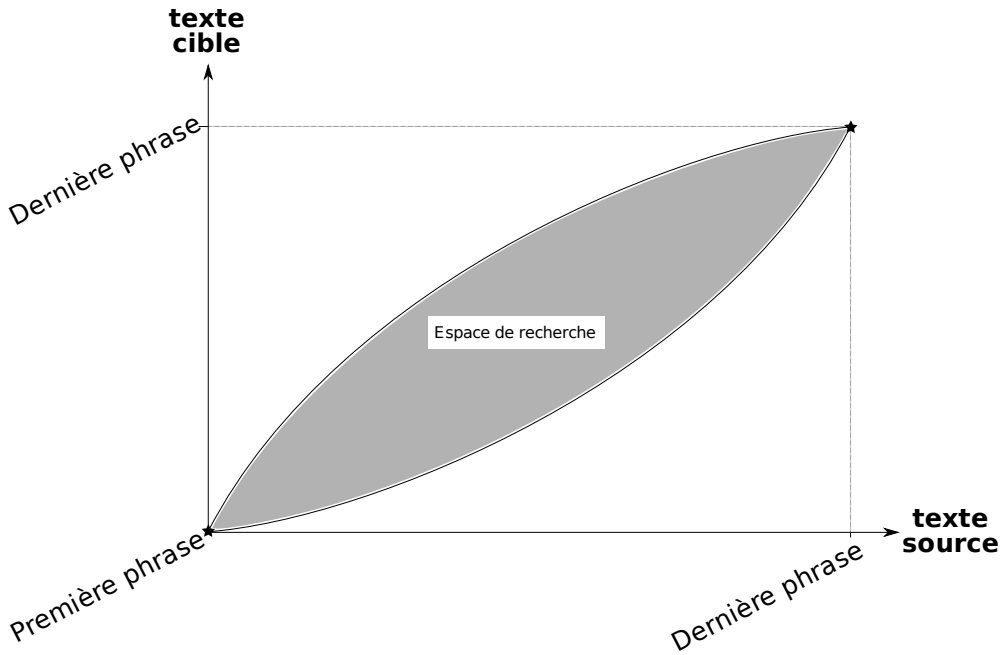
- Alignement basé sur la position et la distribution des mots dans les documents



Corpus parallèles

- Alignement basé sur la position et la distribution des mots dans les documents
- Alignement basé sur des mots déjà connus, utilisés comme points d'ancrage pour aligner leurs voisins

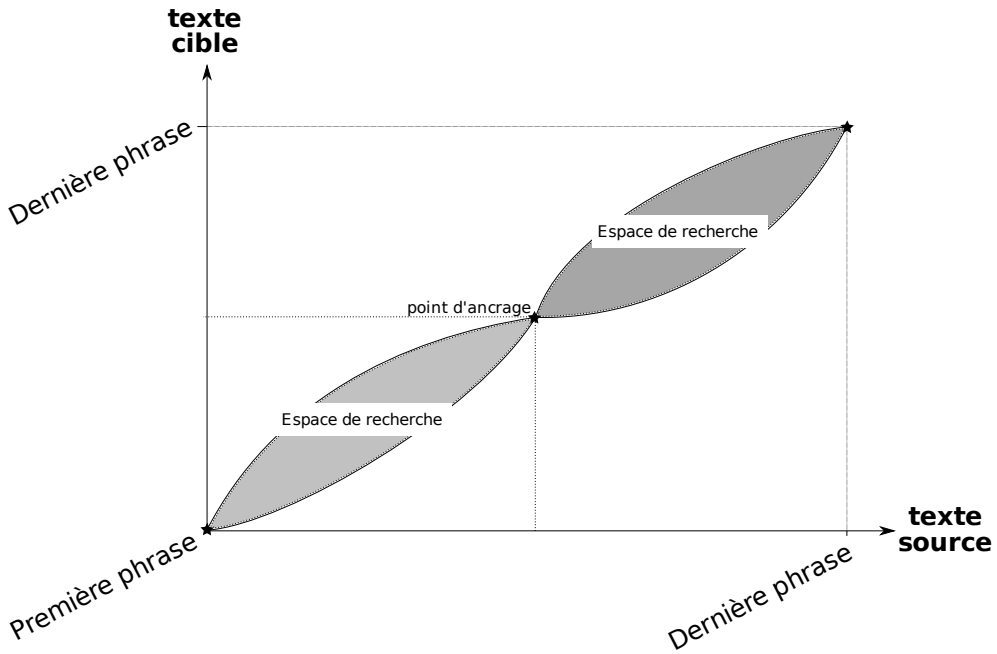
(Véronis, 2000)



Corpus parallèles

- Alignement basé sur la position et la distribution des mots dans les documents
- Alignement basé sur des mots déjà connus, utilisés comme points d'ancrage pour aligner leurs voisins

(Véronis, 2000)



Corpus comparables

- 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

Corpus comparables

- 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.

Corpus comparables

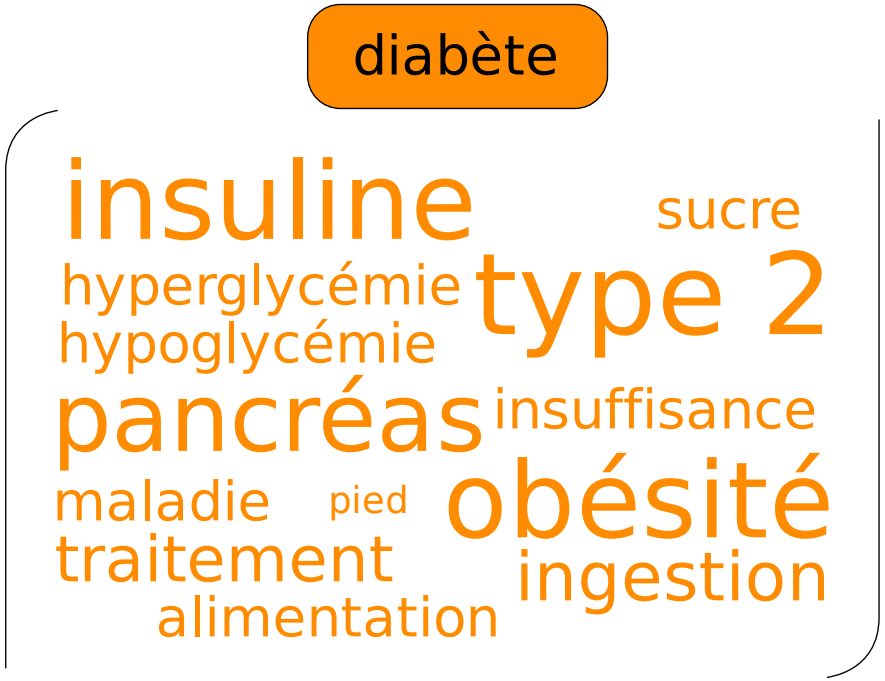
- Rapp (1995) et Fung (1995) introduisent l'alignement à partir de corpus *non-parallèles*
- Tous deux s'appuient sur l'idée de caractériser le *contexte* des mots à traduire, plutôt que des informations sur leurs positions
- Fung (1995) s'appuie sur les bigrammes (hétérogénéité à gauche/à droite), Rapp (1995) s'appuie sur les voisins rencontrés dans une fenêtre de taille fixe autour du mot à traduire.

Firth, 1957

« On reconnaît un mot à ses fréquentations »

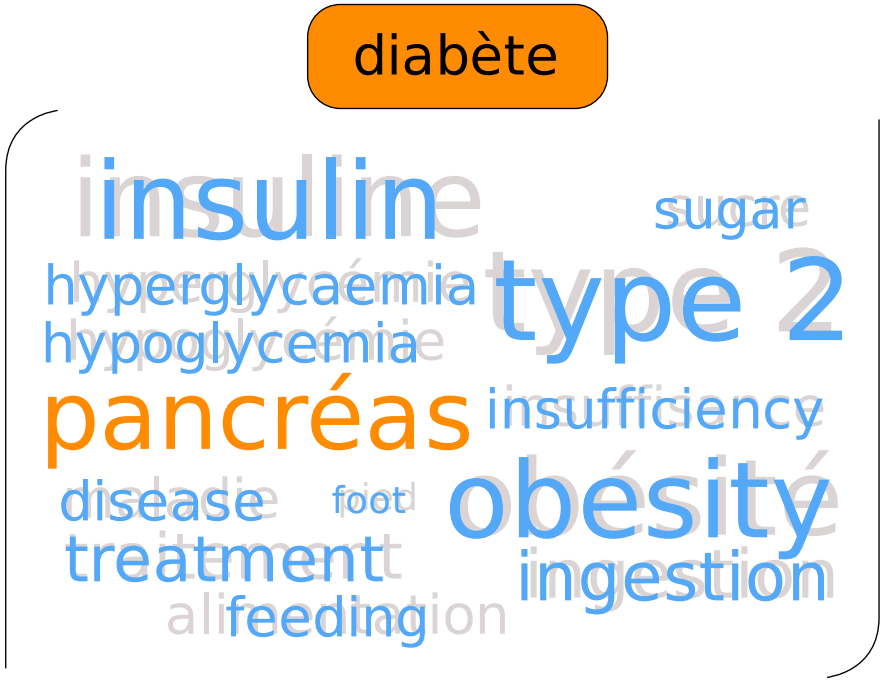
Approche par traduction directe

- **Construction de *vecteurs de contexte***



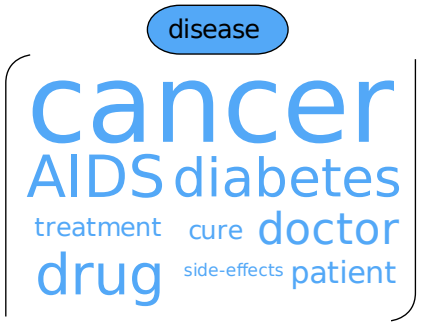
Approche par traduction directe

- **Construction de *vecteurs de contexte***
- Traduction des vecteurs sources vers la langue cible



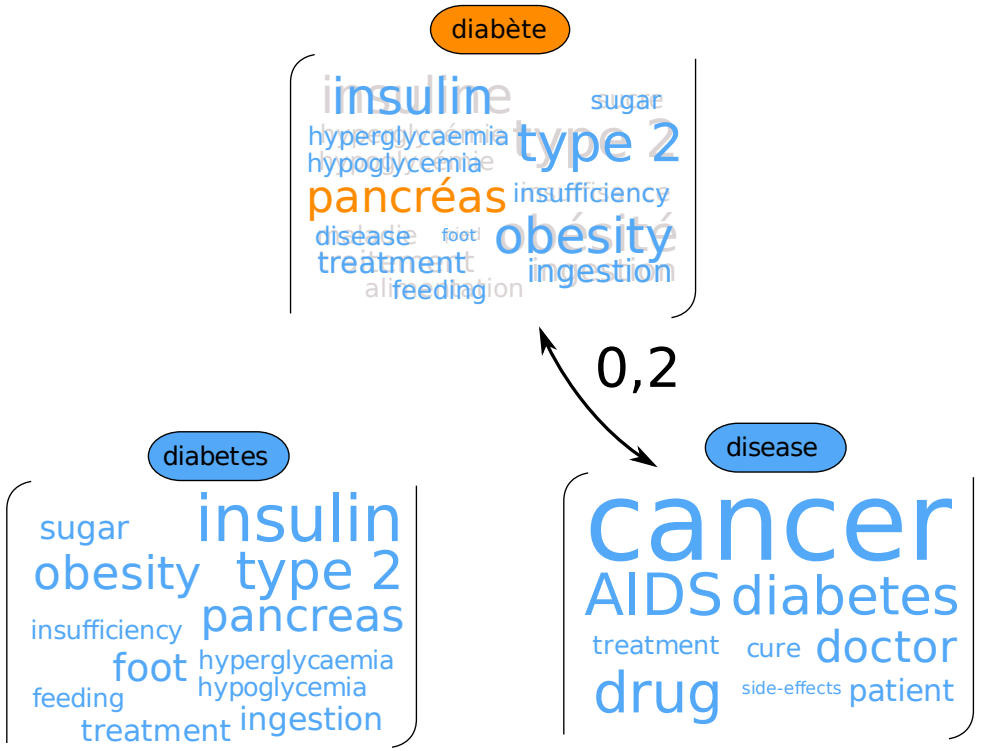
Approche par traduction directe

- **Construction de *vecteurs de contexte***
- Traduction des vecteurs sources vers la langue cible
- Calcul de la similarité entre vecteurs



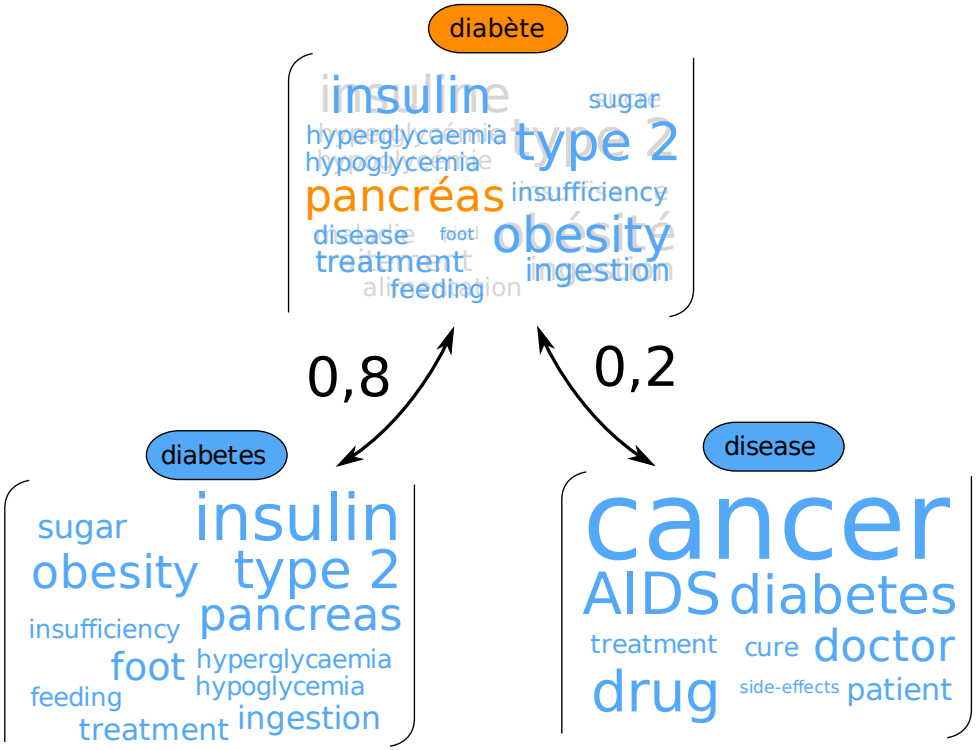
Approche par traduction directe

- **Construction de *vecteurs de contexte***
- Traduction des vecteurs sources vers la langue cible
- Calcul de la similarité entre vecteurs



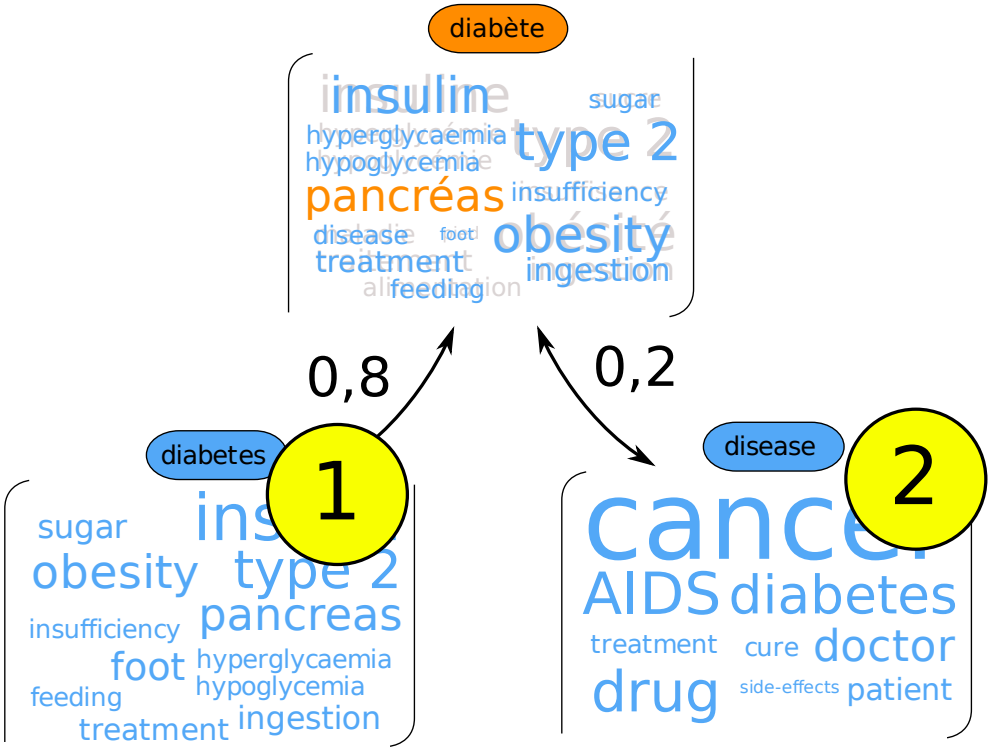
Approche par traduction directe

- **Construction de *vecteurs de contexte***
- Traduction des vecteurs sources vers la langue cible
- Calcul de la similarité entre vecteurs



Approche par traduction directe

- **Construction de *vecteurs de contexte***
- Traduction des vecteurs sources vers la langue cible
- Calcul de la similarité entre vecteurs
- → Liste ordonnée de candidats à la traduction



Emphase : construction des vecteurs de contexte

- Collecte de tous les élements (pertinents) dans une fenêtre donnée autour du mot à caractériser
- Calcul de l'*association* (indépendance statistique) entre la tête du vecteur et ses élements

Emphase : construction des vecteurs de contexte

- Collecte de tous les élements (pertinents) dans une fenêtre donnée autour du mot à caractériser
- Calcul de l'*association* (indépendance statistique) entre la tête du vecteur et ses élements
- Exemple : l'*Information Mutuelle*

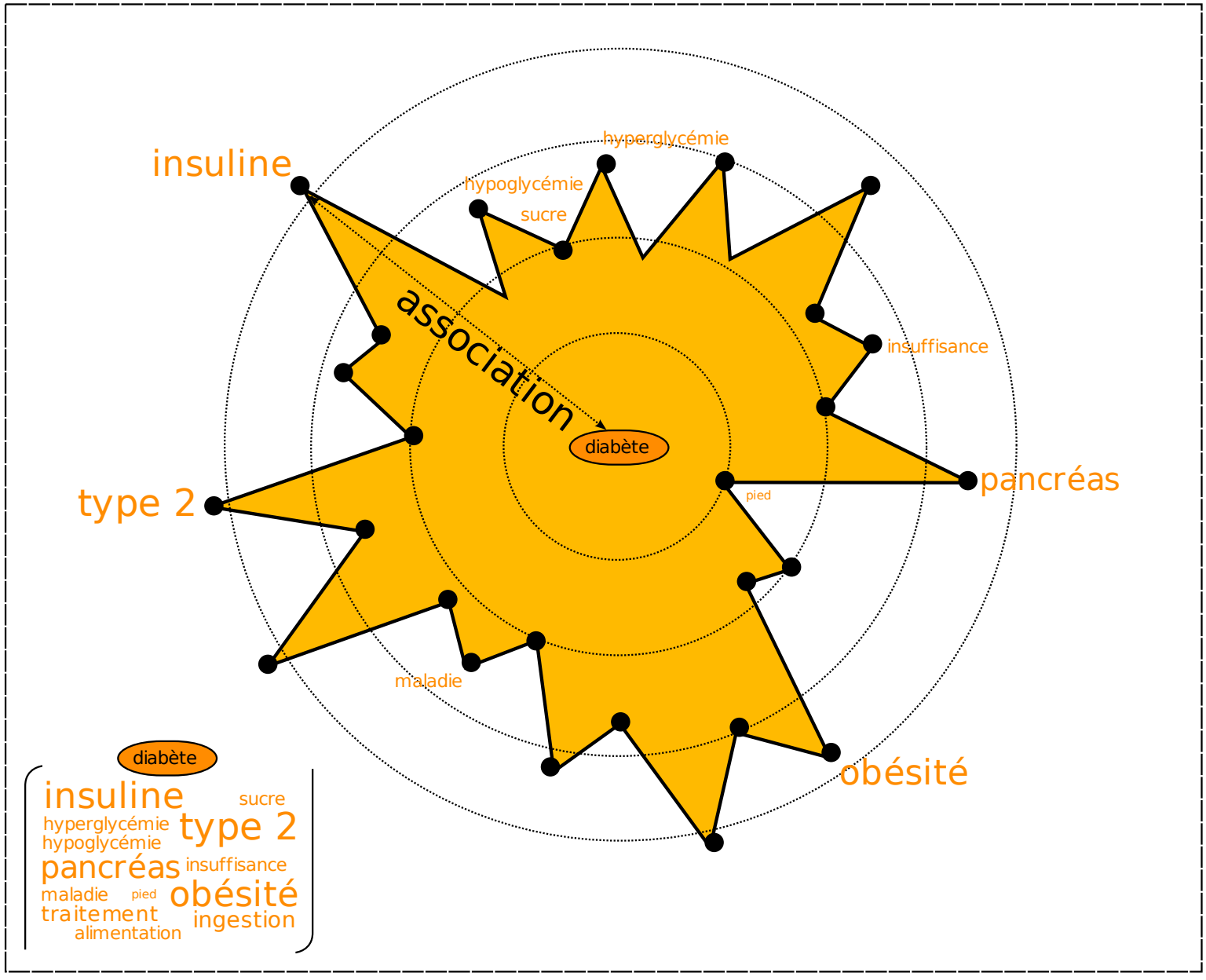
Emphase : construction des vecteurs de contexte

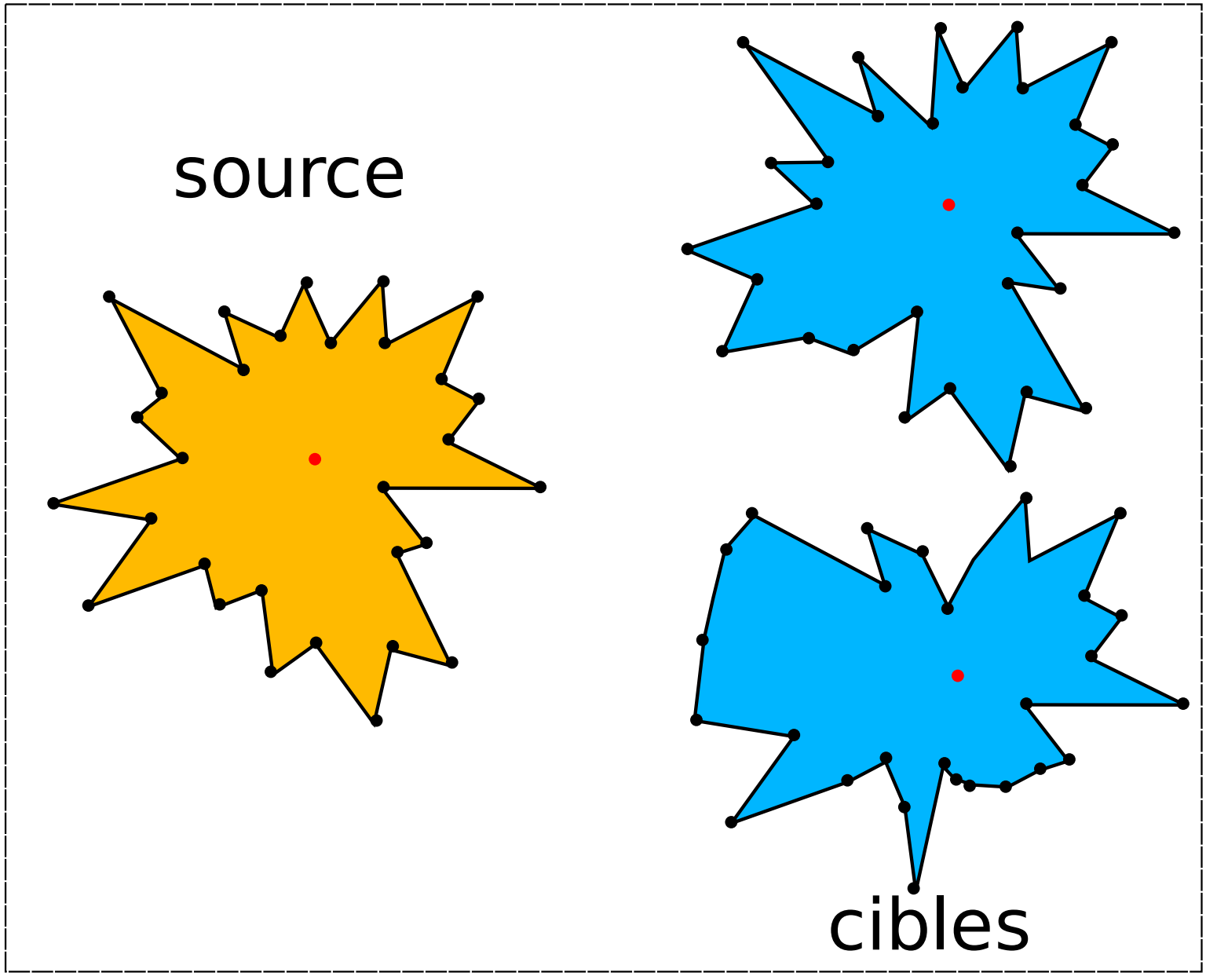
- Collecte de tous les élements (pertinents) dans une fenêtre donnée autour du mot à caractériser
- Calcul de l'*association* (indépendance statistique) entre la tête du vecteur et ses élements
- Exemple : l'*Information Mutuelle*
- $IM = log\frac{O}{E}$

Emphase : construction des vecteurs de contexte

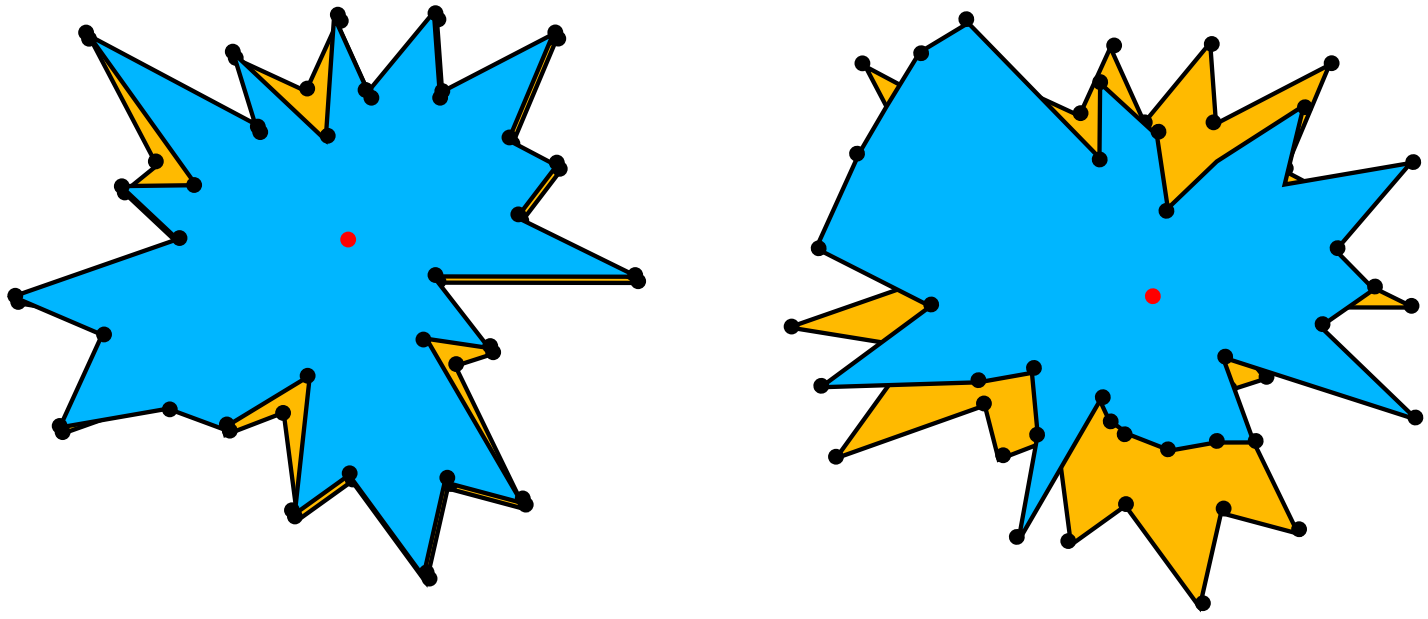
- Collecte de tous les élements (pertinents) dans une fenêtre donnée autour du mot à caractériser
- Calcul de l'*association* (indépendance statistique) entre la tête du vecteur et ses élements
- Exemple : l'*Information Mutuelle*
- $IM = log\frac{O}{E}$

- $\longrightarrow$ Obtention d'un *Motif d'Association*, pour un mot et ses voisins.


insuline
hyperglycémie
hypoglycémie
sucre
association
insuffisance
diabète
pancréas
type 2
pied
maladie
obésité
diabète
insuline
sucre
hyperglycémie
type 2
hypoglycémie
pancréas
insuffisance
maladie
pied
obésité
traitement
ingestion
alimentation


source
cibles

comparaison



Problème

- La construction de motifs significatifs nécessite des jeux de données volumineux

Problème

- La construction de motifs significatifs nécessite des jeux de données volumineux
- Travail sur des textes spécialisés

Problème

- La construction de motifs significatifs nécessite des jeux de données volumineux
- Travail sur des textes spécialisés
- *Faibles volumes de matériaux textuels*

Points d'ancrage

- Idée : compenser le manque de données en s'appuyant sur des *éléments de confiance*

Points d'ancrage

- Idée : compenser le manque de données en s'appuyant sur des *éléments de confiance*
- $\Rightarrow$ Points d'ancrage !

Points d'ancrage

- Idée : compenser le manque de données en s'appuyant sur des *éléments de confiance*
- $\Rightarrow$ Points d'ancrage !
- Conceptuellement assez proche des points d'ancrage dans les corpus parallèles

Points d'ancrage

- Idée : compenser le manque de données en s'appuyant sur des *éléments de confiance*
- $\Rightarrow$ Points d'ancrage !
- Conceptuellement assez proche des points d'ancrage dans les corpus parallèles
  - Exploiter les points d'ancrage pour rendre les vecteurs de contexte plus discriminants
  - $\longrightarrow$ rapprocher les vecteurs traductions
  - $\longrightarrow$ éloigner les vecteurs non traductions

Points d'ancrage

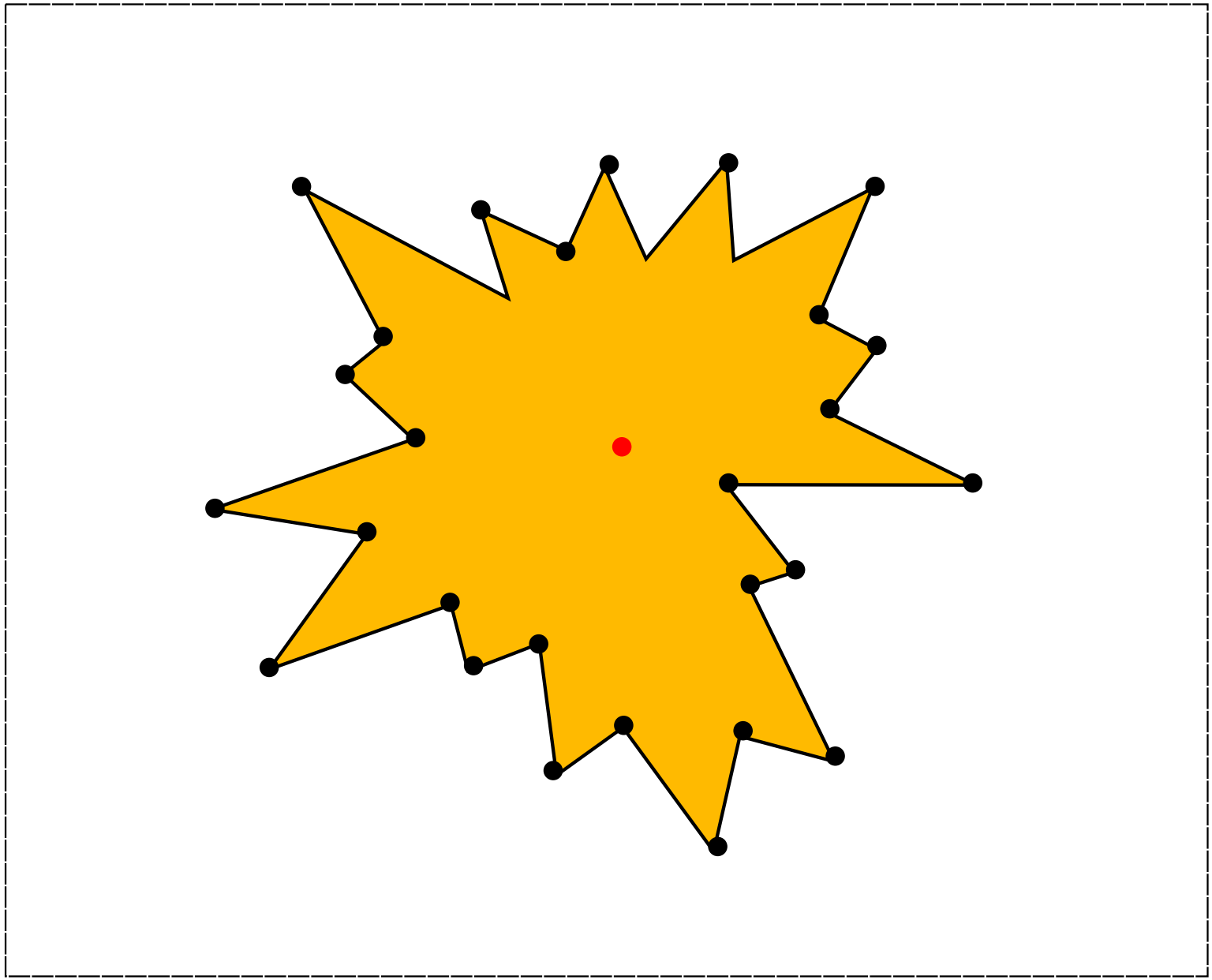
- Idée : compenser le manque de données en s'appuyant sur des *éléments de confiance*
- $\Rightarrow$ Points d'ancrage !
- Conceptuellement assez proche des points d'ancrage dans les corpus parallèles
  - Exploiter les points d'ancrage pour rendre les vecteurs de contexte plus discriminants
  - $\longrightarrow$ rapprocher les vecteurs traductions
  - $\longrightarrow$ éloigner les vecteurs non traductions

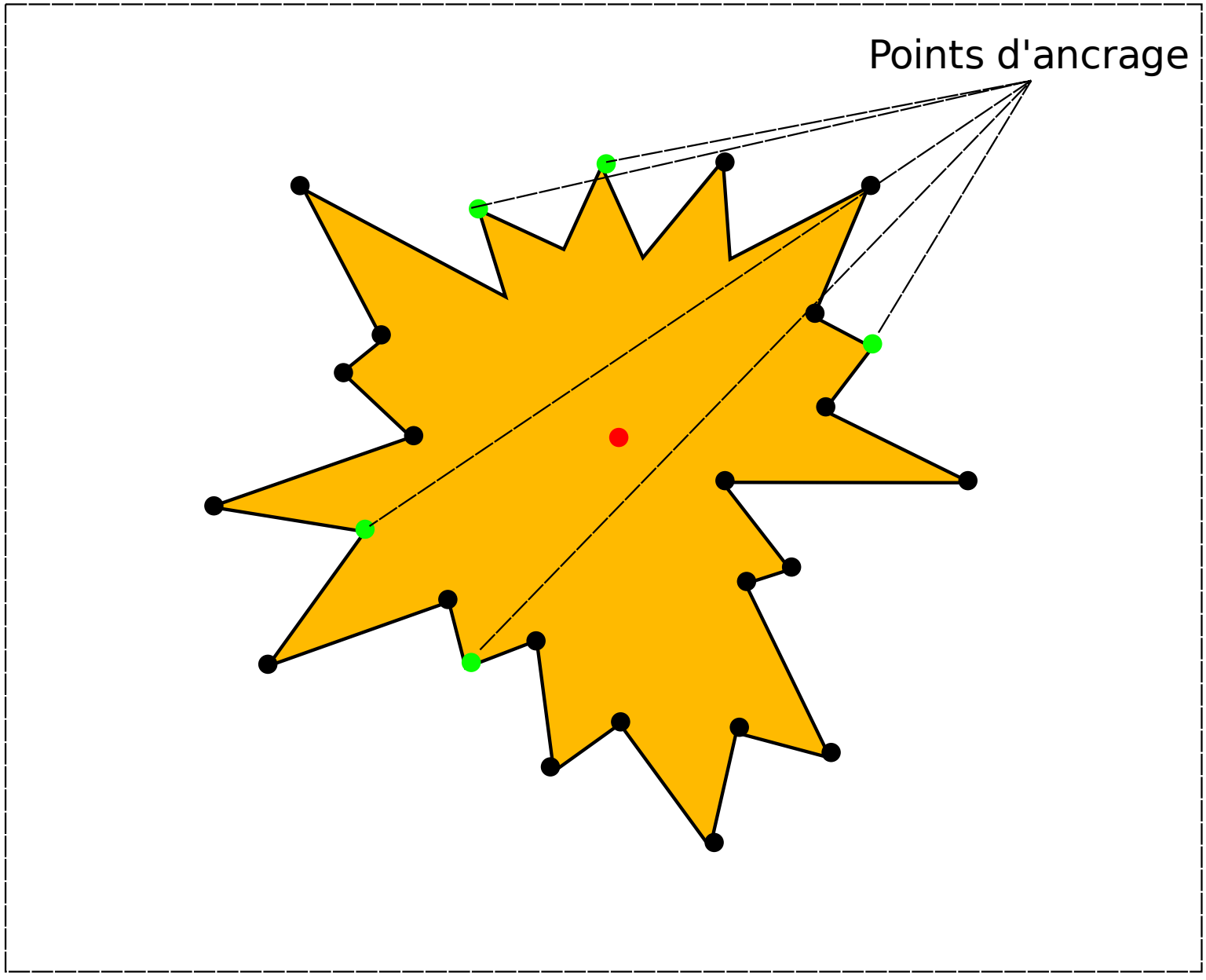
Propriétés

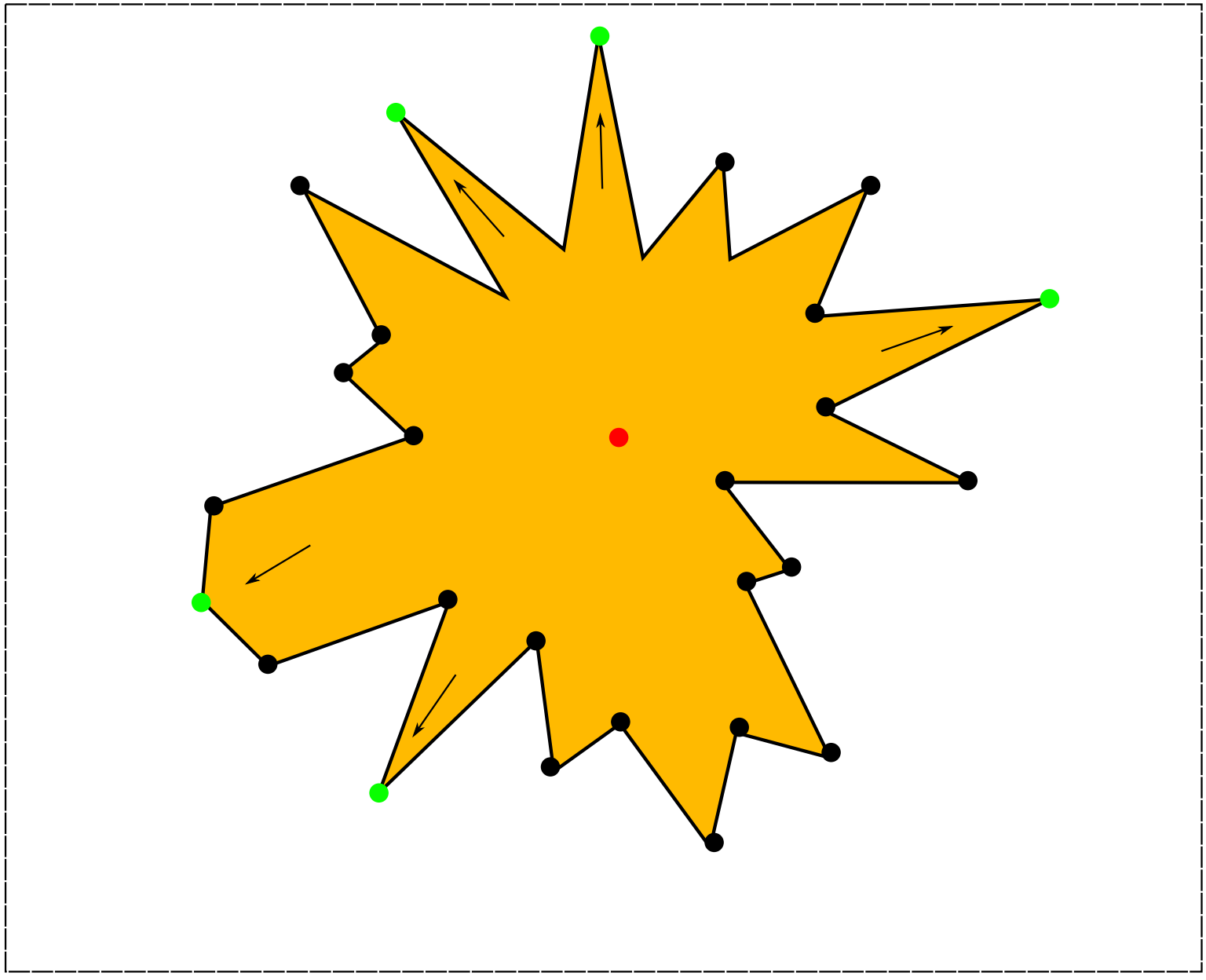
1. Pertinents vis-à-vis des thématiques du corpus
2. Détectables automatiquement
3. Traductions stables

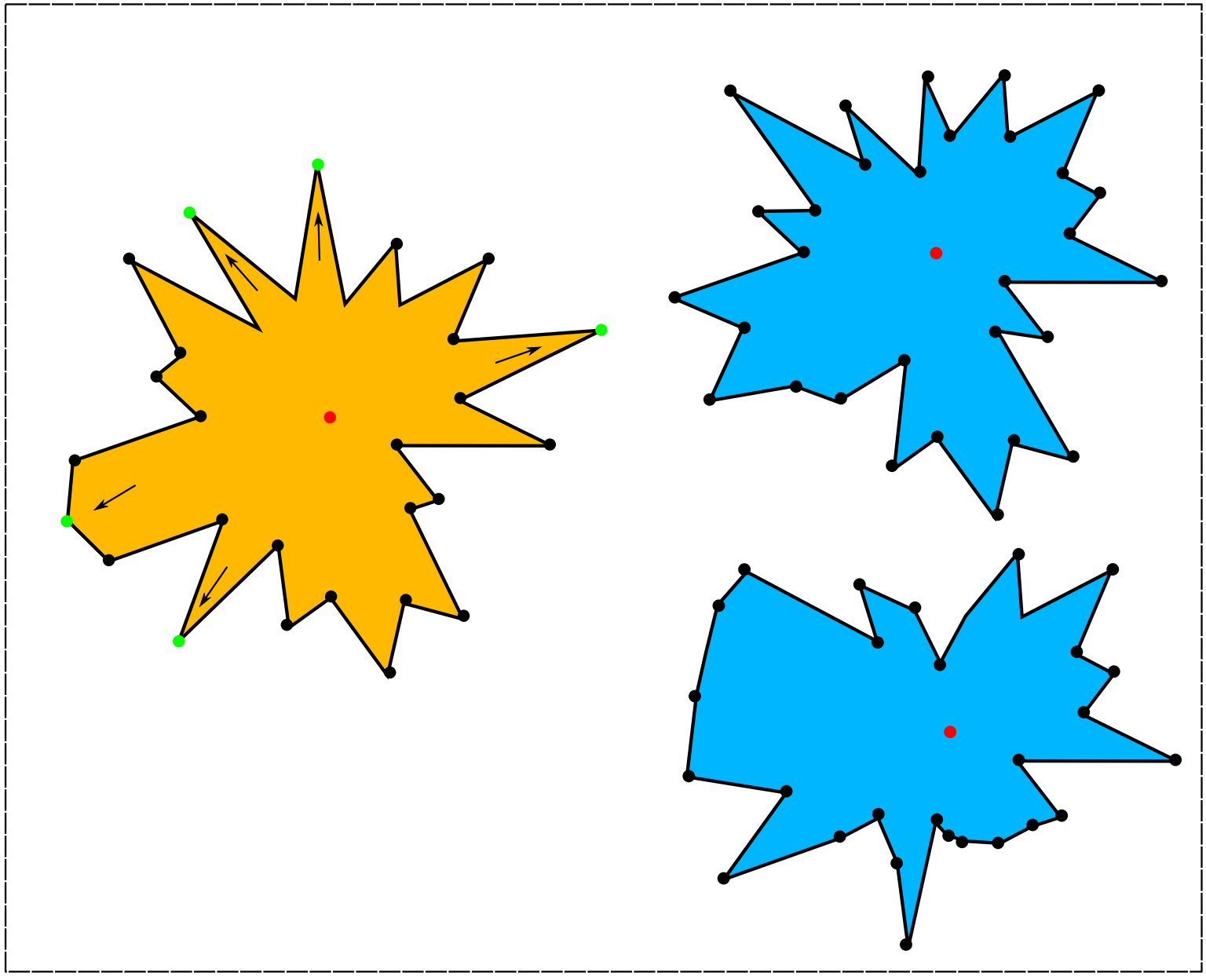
Exploitation des points d'ancrage

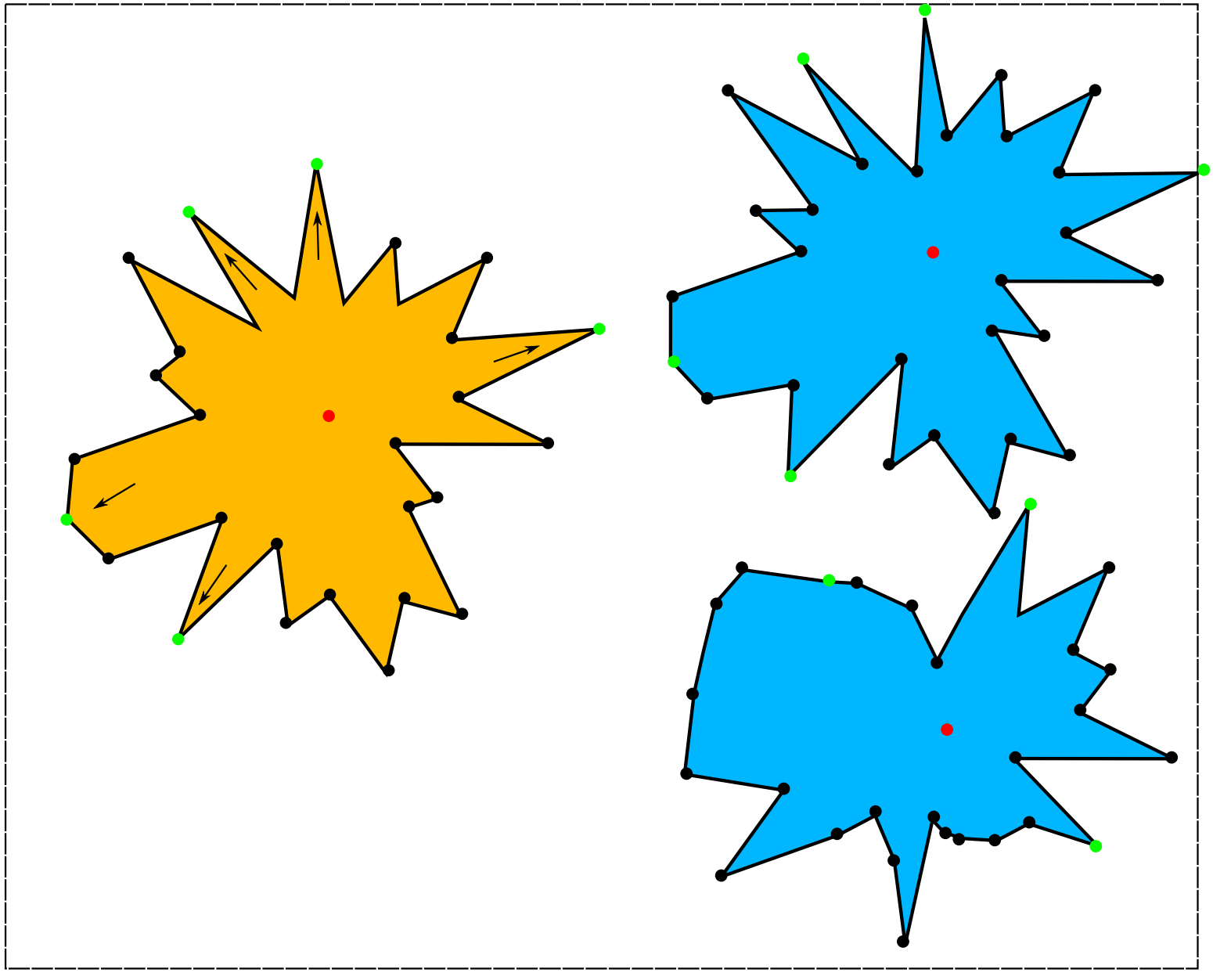
- Idée : construire le motif d'association en priorité sur les points d'ancrage, puis sur les autres éléments
- Augmentation articielle du score d'association des points d'ancrage
- « Déformation » du motif d'association en faveur des points d'ancrage

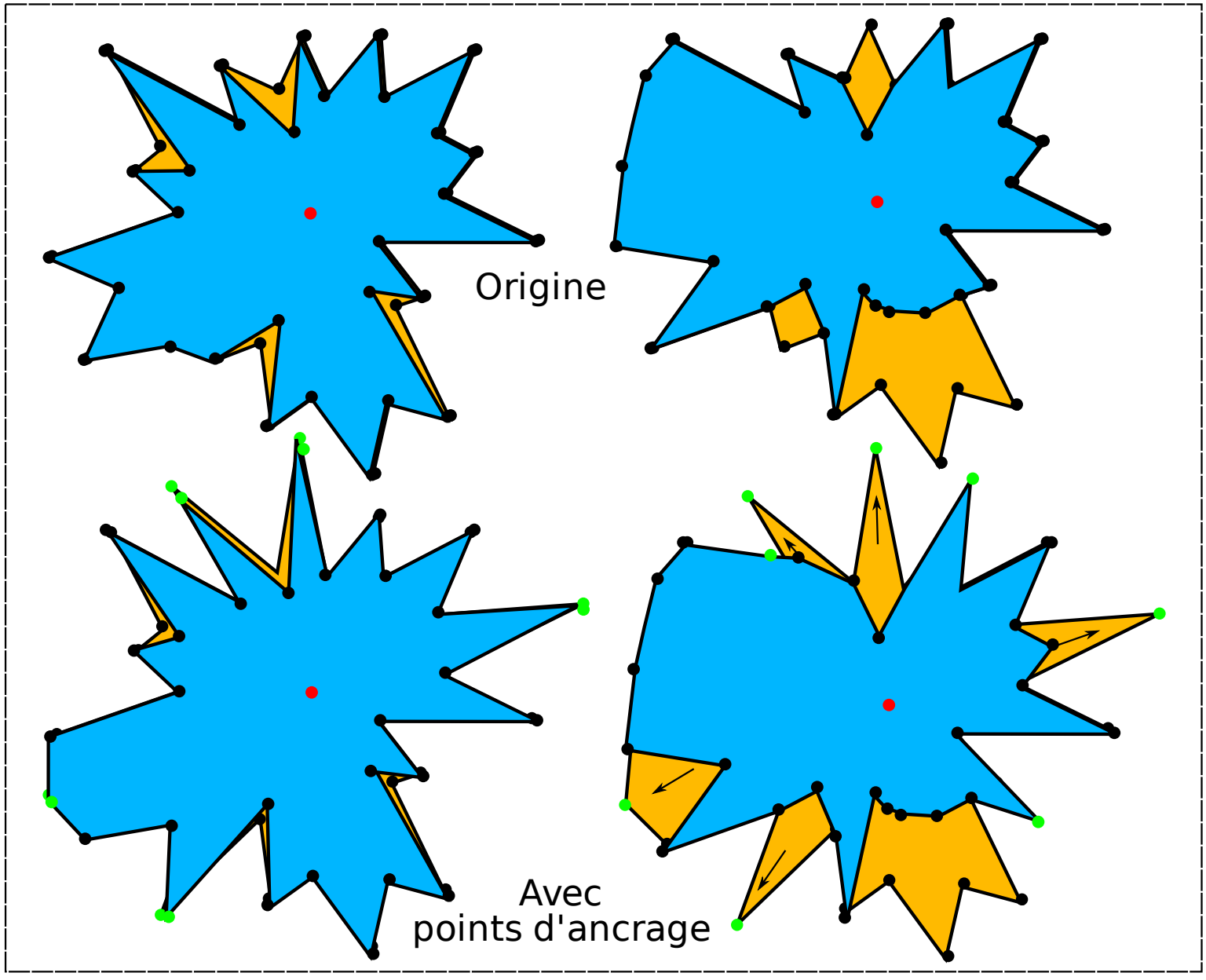



Points d'ancrage








Origine
Avec
points d'ancrage

Contexte

Utilisation d'un corpus comparable anglais, japonais, français

- Domaine *médical*
- Thème *Alimentation et diabète*
- Registre *scientifique*
- Environ 250 000 mots par partie

Points d'ancrage (2)

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

- Translittérations japonaises (et leurs correspondances en français et en anglais)
- Composés savants anglais/français (et leurs traductions en japonais)

Translittérations

- Adaptation phonétique d'un mot aux contraintes du japonais
- Exemple : インスリン/i-n-su-ri-n (insulin/insuline)
- Facile à identifier (syllabaire dédié)
- Alignement automatique sur la base de la prononciation
- Couvre un vocabulaire spécifique, dans le cas des documents *scientifique* (Ito, 2007)
- Emprunt à l'anglais, mais alignement possible avec le français
- Détectées avec un outil dédié à l'alignement anglais/japonais (Tsuji, 2005)

Composés savants

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

Composés savants

- Mots construits sur des racines grecques et latines (Namer, 2005)
- *psychologie*, construit avec le préfixe *psycho-* et le suffixe *-logie*
- Dérivations régulières entre l'anglais et le français (Claveau, 2007)
- *logy* (en) $\longrightarrow$ *logie* (fr)

Composés savants

- Mots construits sur des racines grecques et latines (Namer, 2005)
- *psychologie*, construit avec le préfixe *psycho-* et le suffixe *-logie*
- Dérivations régulières entre l'anglais et le français (Claveau, 2007)
- *logy* (en) $\rightarrow$ *logie* (fr)
- Caractéristique d'un vocabulaire *scientifique*

Composés savants

- Mots construits sur des racines grecques et latines (Namer, 2005)
- *psychologie*, construit avec le préfixe *psycho-* et le suffixe *-logie*
- Dérivations régulières entre l'anglais et le français (Claveau, 2007)
- *logy* (en) $\rightarrow$ *logie* (fr)
- Caractéristique d'un vocabulaire *scientifique*
- Détectés à l'aide d'une liste d'affixes

Protocole

Trois expériences

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

- Points d'ancrage *translittérations* : 589 (en/jp) 526 (fr/jp)
- Points d'ancrage *composés savants* : 604 (en/jp) 819 (fr/jp)

- Utilisation d'une liste de référence de 98 termes
- Taille de fenêtre : 25

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 ($Top > 50$)

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

Conclusion, discussion

- Nouvelle hypothèse pour l'alignement de lexique à partir de corpus comparables spécialisés
- Résultats globalement significatifs
- Hypothèse extensible à d'autres types de vocabulaire

Conclusion, discussion

- Nouvelle hypothèse pour l'alignement de lexique à partir de corpus comparables spécialisés
- Résultats globalement significatifs
- Hypothèse extensible à d'autres types de vocabulaire

- Expérience à reproduire avec d'autres points d'ancrage
- À définir en fonction des couples de langues impliquées
- Utilisation d'autres techniques transversales en TALN
- (détection des cognats, techniques de RI. . .)

Fin

Merci de votre attention

Prise en compte des points d'ancrage

$$assoc_{(PA)}{}_a^v = assoc_a^v + \beta \quad (1)$$

Influence des points d'ancrage