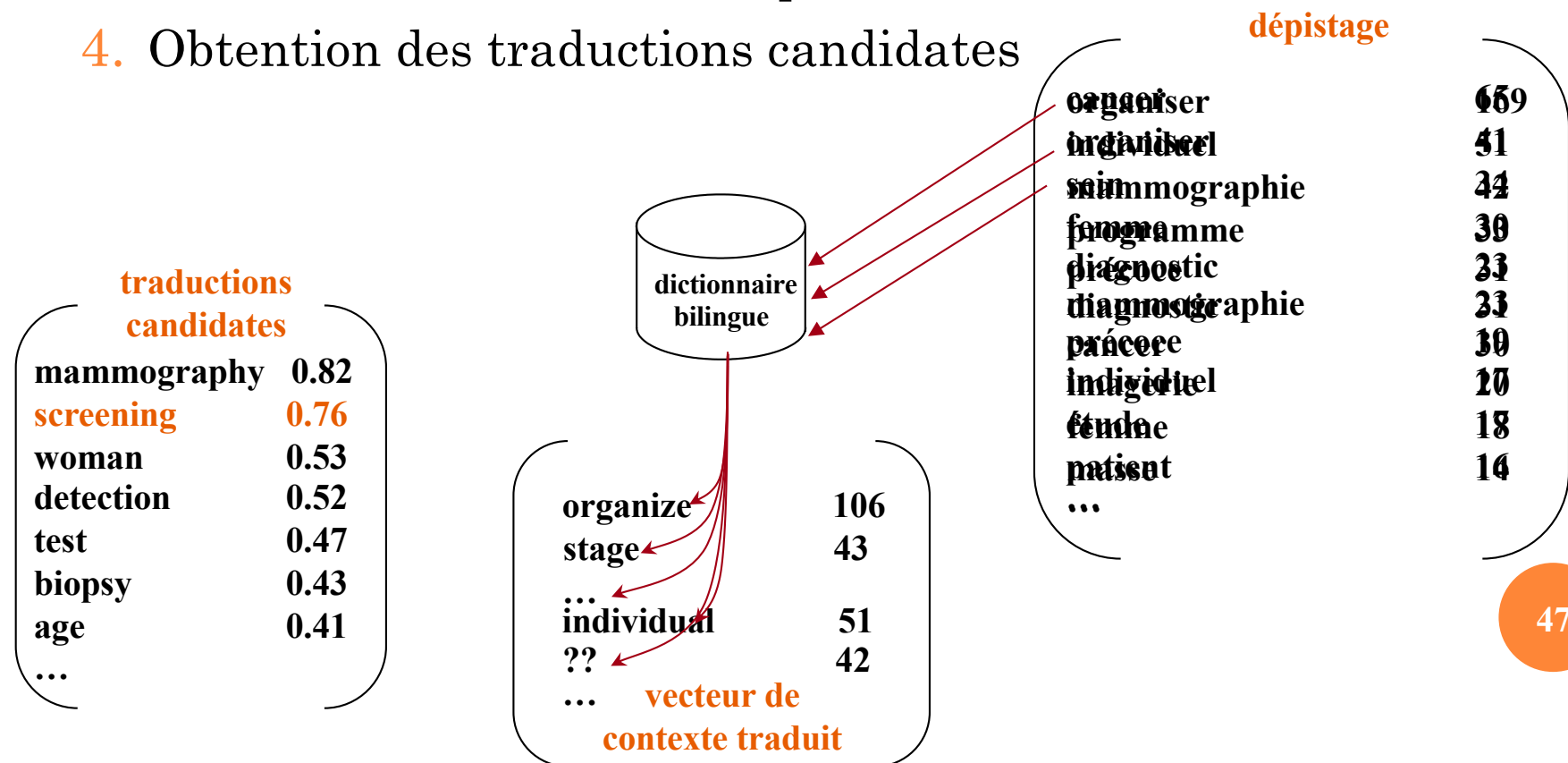


Méthodes d'alignement

- Méthodes d'extraction de lexiques bilingues à partir de corpus comparables :
 - basées sur une analyse du contexte lexical des mots et reposent sur la simple observation qu'un mot et sa traduction tendent à apparaître dans les mêmes contextes lexicaux
 - trouvent un ancrage dans la proposition de Firth (1957) : « *You shall know a word by the company it keeps* »
- Mise en œuvre qui repose sur l'identification d'affinités du premier ordre : « *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word* » (Grefenstette, 1994a, p. 279) : méthode directe ou méthode standard

Méthode standard

- Méthode standard (Fung & McKeown, 1997; Rapp, 1999) :
 1. Identification des contextes lexicaux
 2. Transfert d'une unité à traduire
 3. Identification des vecteurs proches de l'unité à traduire
 4. Obtention des traductions candidates



Méthode standard

- Pour un corpus comparable Français/Japonais de 1,5 million de mots relatif au diabète
 - pour une liste de 100 termes simples à traduire (dont la traduction est un terme simple) : TOP₁₀ : 51% et TOP₂₀ : 60%

Référence	Domaine/Langues/Taille	TOP ₁₀	TOP ₂₀
Déjean & Gaussier (2002)	Médical/Allemand-Anglais/0.1	43	51
Déjean & Gaussier (2002)	Général/Allemand-Anglais/8	79	84
Chiao & Zweigenbaum (2002)	Médical/Français-Anglais/1.2	61	94

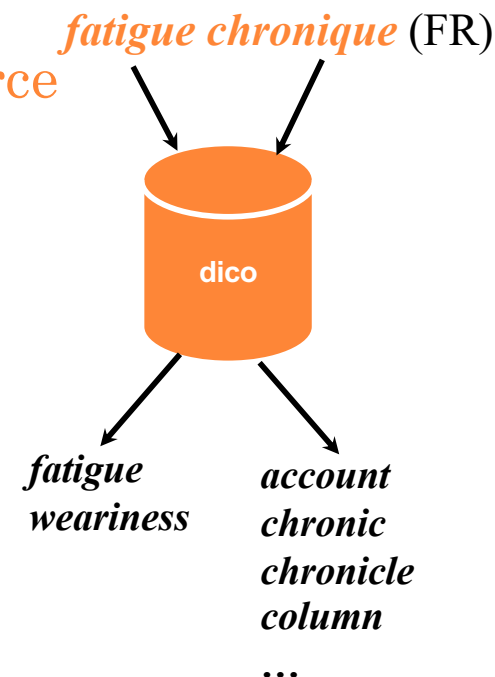
- pour une liste de 60 termes simples et complexes à traduire (dont la traduction d'un terme simple ne peut être un terme simple) : TOP₁₀ : 17% et TOP₂₀ : 20%

Autres méthodes

- Quelques améliorations de l'approche directe :
 - Chiao (2004) introduit une hypothèse de symétrie distributionnelle
 - Prochasson et Morin (2009) introduisent la notion de point d'ancrage (translitérations et composés savants)
 - Autres méthodes :
 - méthode par similarité interlangue et ses variantes (Déjean et Gaussier 2002, Daille et Morin, 2005; Hazem et al. 2009)
 - Méthode numérique CCA, ICA, PLSA... (Gaussier et al, 2004; Hazem et Morin 2012)
 - ...
- ⇒ Ces méthodes semblent avoir atteint leur limite en termes de performance

APPROCHE COMPOSITIONNELLE

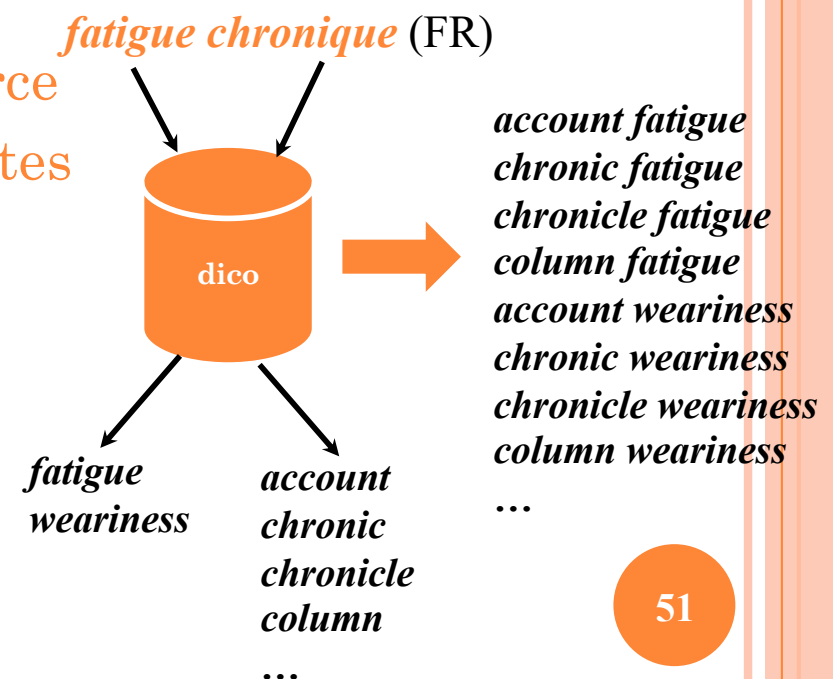
- Compositionnalité : « *the meaning of the whole is a function of the meaning of the parts* » (Keenan and Faltz, 1985, p. 24-25)
- Traduction compositionnelle à partir de corpus comparables repose sur les étapes suivantes (Grefenstette, 1999; Tanaka, 2002; Robitaille et al., 2006) :
 - Traduction du terme en langue source
 - chaque composant du terme de la langue source est traduit en langue cible



APPROCHE COMPOSITIONNELLE

- Compositionnalité : « *the meaning of the whole is a function of the meaning of the parts* » (Keenan and Faltz, 1985, p. 24-25)
- Traduction compositionnelle à partir de corpus comparables repose sur les étapes suivantes (Grefenstette, 1999; Tanaka, 2002; Robitaille et al., 2006) :

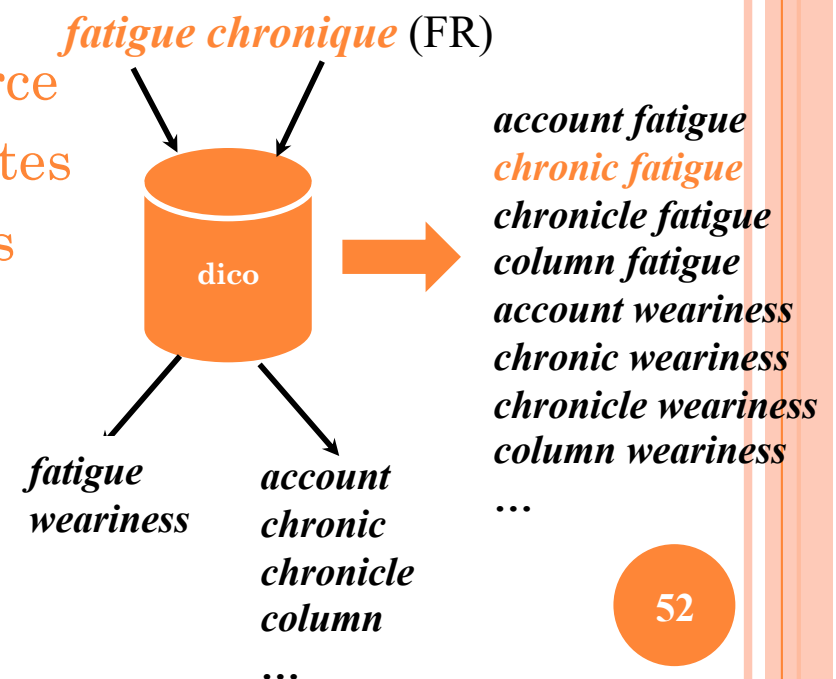
- Traduction du terme en langue source
- Génération des traductions candidates
 - l'ensemble des combinaisons sont générées sans tenir compte de l'ordre des composants



APPROCHE COMPOSITIONNELLE

- Compositionnalité : « *the meaning of the whole is a function of the meaning of the parts* » (Keenan and Faltz, 1985, p. 24-25)
- Traduction compositionnelle à partir de corpus comparables repose sur les étapes suivantes (Grefenstette, 1999; Tanaka, 2002; Robitaille et al., 2006) :

- Traduction du terme en langue source
- Génération des traductions candidates
- Sélection des traductions candidates
 - les traductions candidates sont les termes identifiés en langue cible ordonnés suivant leur nombre d'occurrences



APPROCHE COMPOSITIONNELLE

- Méthode de repli de Robitaille et al. (2006) :
 - Décomposer un terme à traduire de longueur n (avec $n > 2$ mots pleins) en toutes les combinaisons de termes de longueur inférieure ou égale à n
technique du ganglion sentinelle →
 - [technique du ganglion sentinelle]
 - [technique du ganglion] [sentinelle]
 - [technique] [ganglion sentinelle]
 - [technique] [ganglion] [sentinelle]

Approche compositionnelle basée sur la morphologie

- Hypothèse :

- La morphologie dérivationnelle est un processus compositionnel au niveau de la traduction

- Patrons liés par dérivation :

- $N_1 \text{ ADJ} \rightarrow N_1 \text{ Prep Art? } N_2$ avec $M(\text{ADJ}, N_2)$
index glycémique → index de la glycémie
- $N_1 \text{ Prep Art? } N_2 \rightarrow N_2 \text{ ADJ}$ avec $M(N_1, \text{ADJ})$
gravité du risque → risque grave
- $N_1 \text{ ADJ} \rightarrow N_2 \text{ ADJ}$ avec $M(N_1, N_2)$
entreposage frigorifique → entrepôt frigorifique
- $N \text{ ADJ}_1 \rightarrow N \text{ ADJ}_2$ avec $M(\text{ADJ}_1, \text{ADJ}_2)$
phénol polymérisé → phénol non-polymérisé

Approche compositionnelle basée sur la morphologie

- En français, il y a deux catégories principales d'adjectifs :
 - épithète: *important*
 - relationnel : *laitier*
- Adjectifs qui peuvent être argument d'un nom prédicatif :
 - *production laitière*
 - *production importante*
- Adjectifs dénominaux :
 - *lait* → *laitier*
 - *gaz* → *gazeux*
- Suffixes appropriés :
 - *-ain, -aire, -al, -el, -estre, -ien, -ier, -il(e), -in, -ique*

Approche compositionnelle basée sur la morphologie

- Synonymie entre adjectif dénominal et syntagme prépositionnel :
 - *forestier* \leftrightarrow *de la forêt*
- Règles de recodage :
 - (source) $N_1 \text{ ADJ} \rightarrow N_1 \text{ Prep Art? } N_2 \Leftrightarrow$ (cible) $N_2 N_1$
 - $N_1 \text{ ADJ} \rightarrow N_1 \text{ Prep Art? } N_2 \quad M(\text{ADJ}, N_2)$
 - $M(\text{ADJ}, N_2) = [-ique, -ie]$
 - $M(\text{ADJ}, N_2) = [-ulaire, -le]$
 - $M(\text{ADJ}, N_2) = [-eux,]$
 - ...
 - (French) *essence / N1 forestière / ADJ* \rightarrow *essence / N1 de / Prep la / Art forêt / N2* \Leftrightarrow (English) *tree / N2 species / N1*

Approche compositionnelle basée sur la morphologie

- Corpus comparable Français/Japonais :
 - [mixed corpora] 1.5 million de mots rassemblant des documents scientifiques et vulgarisés relatifs au cancer du sein
- Deux listes de référence composées de termes de structure N Adj Français :
 - [N ADJE] 749 MWTs
 - [N ADJR] 829 MWTs

Approche compositionnelle basée sur la morphologie

○ Approche compositionnelle

	# MWT FR	# MWT JP	# Correct JP
[N ADJE]	76	98	68
[N ADJR]	8	8	5

○ Approche compositionnelle basée sur la morphologie

	# MWT FR	# MWT JP	# Correct JP
[N ADJR]	128	170	150

- (Fr) *traitement hormonal* → *traitement aux hormones*
⇔ (Jp) ホルモン療法
- (Fr) *patient diabétique* → *patient de diabète*
⇔ (Jp) 糖尿病患者

Approche compositionnelle basée sur la morphologie

- À partir de 859 MWT français de structure N ADJR à traduire en japonais :
 - 30 termes (5.1%) par le dictionnaire
 - 5 termes (0.6%) par la méthode compositionnelle
 - 150 termes (17.5%) par la méthode compositionnelle basée sur la morphologie

APPROCHE COMPOSITIONNELLE

- Approche facile à mettre en oeuvre mais qui échoue quand :
 1. au moins un composant du terme à traduire n'est pas présent dans le dictionnaire bilingue
 - ce composant ne peut être traduit
 2. la traduction candidate est valide mais n'a pas été identifiée par le programme d'extraction de terminologie en langue cible :
 - la traduction candidate n'est pas présente dans la partie cible du corpus comparable
 - le concept existe dans la langue cible mais sous une forme non terminologique
 - une erreur durant les prétraitements a induit la non reconnaissance du terme
 3. la traduction candidate n'est pas valide
 - éventuellement en raison d'un problème de non-compositionnalité, fertilité, variante

COMPOSITIONAL METHOD WITH CONTEXT-BASED PROJECTION (CMCBP)

- A first solution would be to find synonyms in the source language:
 - For low-frequency words, Pekar et al. (2006) predicted missing co-occurrence values based on similar words in the same language
 - For difficult translations, Sharoff et al. (2009) find similar words in the source language to produce a more reliable similarity
 - We already perform a subset of the synonymy clustering by using a set of term variants instead of one term alone
- ⇒ Principe: use of the context of the words (which are parts of the MWT to be translated) when the compositional approach fails


COMPOSITIONAL METHOD WITH CONTEXT-BASED PROJECTION (CMCBP)

■ CMCBP uses four steps:

1. Computing the context of the MWT

- when a component is not found in the bilingual dictionary, we replace it by co-occurrence information

antécédent familial



familial	322,9
personnel	73,0
cancer	68,1
sein	48,0
mastopathie	38,9
degré	22,6
patient	19,3
mastodynie	17,6
saignement	16,0
...	

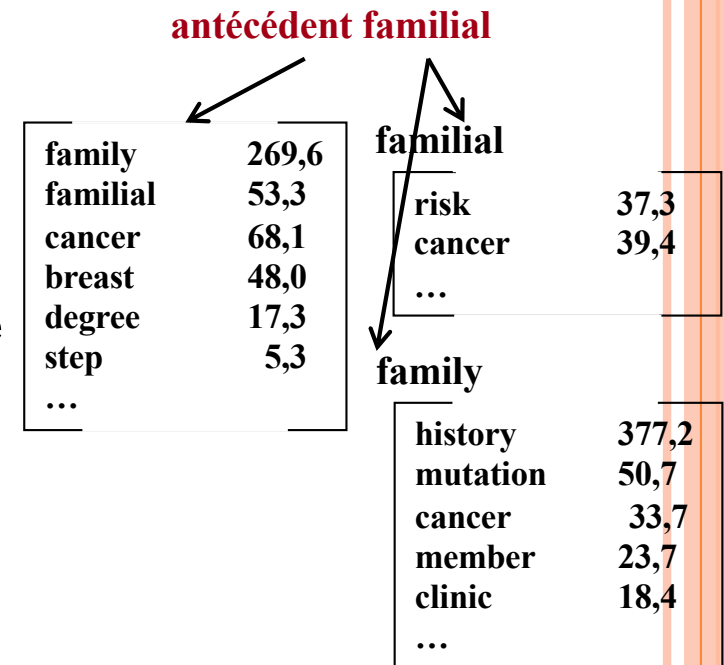
COMPOSITIONAL METHOD WITH CONTEXT-BASED PROJECTION (CMCBP)

■ CMCBP uses four steps:

1. Computing the context of the MWT

2. Transfer of the MWT

- If the component is not found in the dictionary, we translate each element of its context vector
- If the component is found in the dictionary, we use the context vector in the target language of the translation



COMPOSITIONAL METHOD WITH CONTEXT-BASED PROJECTION (CMCBP)

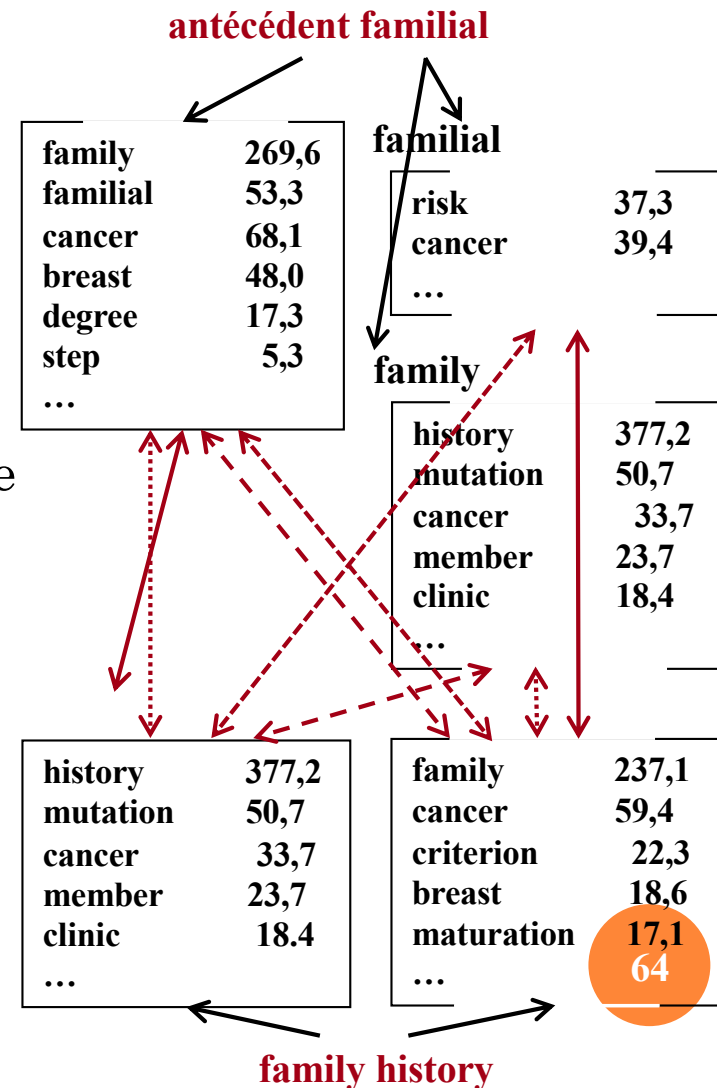
■ CMCBP uses four steps:

1. Computing the context of the MWT

2. Transfer of the MWT

3. Generation of candidate translations

- each MWT of the target language is compared to the transferred MWT through a similarity measure such as Cosine or Weighted Jaccard



COMPOSITIONAL METHOD WITH CONTEXT-BASED PROJECTION (CMCBP)

■ CMCBP uses four steps:

1. Computing the context of the MWT

2. Transfer of the MWT

3. Generation of candidate translations

4. Ranking of candidate translations

- rank the candidate translations in decreasing order of their combination score

antécédent familial



family history	0,75
cancer family	0,57
family remember	0,22
high-risk family	0,18
familial risk	0,06
...	

RESOURCES

○ Comparable corpora

- English/French/German specialized comparable corpus from the medical domain within the sub-domain of ‘breast cancer’
- Articles collected from scientific papers websites for which the title or the keywords of the articles contain *breast cancer* in English, *cancer du sein* in French and *brustkrebs* in German
- ⇒ 118 documents in English, 130 in French and 103 in German (about 530,000 words for English and French languages and 220,000 words for German language)

○ Bilingual dictionary

- French/English dictionary ELRA-M0033 (243,539 translations)
- French/German dictionary ELRA-M0034 (170,967 translations)
- ⇒ Two general language dictionaries which contain only a few terms related to the medical domain

RESOURCES

- Multi-Word Term Test Set :
 - MWT are extracted monolingually through TTC TermSuite (Rocheteau and Daille, 2011)
 - TermSuite first normalises the texts through the linguistic pre-processing steps: tokenisation, part-of-speech tagging and lemmatisation using TreeTagger (Schmid, 1995)
 - TermSuite extracts MWTs whose syntactic patterns correspond either to a canonical or a variation structure
 - The patterns are expressed using MULTEXT part-of-speech tags and are provided for each language
 - The variants handled for MWTs are graphical, morphological, and syntactic

RESOURCES

- Multi-Word Term Test Set :
 - For the French MWT *cancer du sein* (*breast cancer*)
 - base form of N Sp N pattern: *cancer du sein* (*breast cancer*)
 - inflexional variant: *cancers du sein* (*breast cancers*)
 - syntactic variant: *cancer primitif du sein* (*primary breast cancer*)
cancer des ovaires et du sein (*ovarian and breast cancer*)
 - We selected the French MWTs extracted by TermSuite for which the number of occurrences is greater than or equal to 5
- ⇒ The test set is composed of 976 French MWTs (for which 90% of the base forms are only composed of two content words)

EXPERIMENTS

○ Dictionary Look-up

- From the 976 French MWTs to be translated :
 - 51 are recorded in the French/English dictionary
 - 12 in the French/German dictionary
- ⇒ We were unable to generate any translations for 836 French MWTs in English and for 964 French MWTs in German

○ Compositional method

	# trans.	<i>Top₁</i>	<i>Top₅</i>
French/English (836)	140	73.2%	79.1%
French/German (964)	87	88.8%	95.7%

- ⇒ We were unable to generate any translations for 785 French MWTs in English and for 877 French MWTs in German

EXPERIMENTS

- Compositional Method with Context-Based Projection
 - The parameters required for our approach are as follows:
 - size of the context window w is up to 3 (i.e. a seven-word window)
 - association measure is Mutual Information
 - distance measure is Cosine

	# trans.	<i>Top₁</i>	<i>Top₅</i>	<i>Top₁₀</i>	<i>Top₂₀</i>
French/English (836)	514	42.1%	55.4%	56.8%	57.1%
French/German (964)	510	44.3 %	49.4 %	51.2%	51.2%

- These results indicate that the majority of the correct translated MWTs are in fact obtained from the Top 5
- Moreover, the CMCBP retains the advantages of the compositional method. All translations obtained with the compositional method are found in the same rank with the CMCBP

EXPERIMENTS

- From the MWTs correctly translated:
 - A large majority of French MWTs involving a relational adjective:
 - *dépistage*_{DICO} *mammographique*_{CV} (FR) → *mammographic*_{CV} *screening*_{DICO} (EN) *Top*₃
 - Some MWTs with a compositional structure for which one element is not found in the dictionary:
 - *amélioration*_{CV} *significative*_{DICO} (FR) → *significant*_{DICO} *benefit*_{CV} (EN) *Top*₁
 - MWTs without compositional structure:
 - *curage*_{CV} *axillaire*_{CV} (FR) → *axillary*_{CV} *dissection*_{CV} (EN) *Top*₁₁

EXPERIMENTS

- From the MWTs incorrectly translated:
 - English MWTs semantically close to the French MWTs to be translated:
 - *retrospective*_{CV} *study*_{DICO} for *étude*_{DICO} *comparative*_{CV} (*comparative study*)
 - Only a sub-part of the English MWTs is found:
 - *node*_{DICO} *dissection*_{CV} for *curage*_{DICO} *ganglionnaire* (*lymph node dissection*)

SOMMAIRE

1. Introduction
2. Comparabilité des corpus comparables
3. Méthodes d'alignement
4. **Exploitation des résultats d'alignement**
5. Conclusion

Exploitation des résultats d'alignement

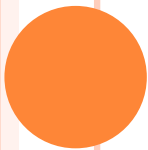
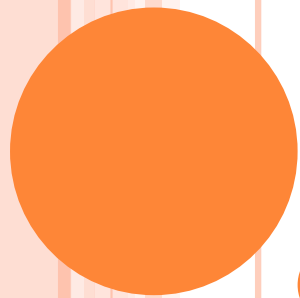
- Les résultats obtenus pour des corpus comparables spécialisés sont moins bons que ceux obtenus :
 - pour des corpus comparables de langue générale
 - pour des corpus parallèles
- Néanmoins, ils permettent d'améliorer les résultats des systèmes de recherche d'information interlingue (Bo, 2012)
- En aide à la traduction, les résultats sont plus contrastés (Delpech, 2011) :
 - liste trop bruitée pour un traducteur
 - liste à plat trop peu informative pour un traducteur

SOMMAIRE

1. Introduction
2. Comparabilité des corpus comparables
3. Méthodes d'alignement
4. Exploitation des résultats d'alignement
5. **Conclusion**

CONCLUSION

- L'extraction de lexiques bilingues à partir de corpus comparables est encore un domaine relativement neuf
- Trois stratégies actuellement :
 - Améliorer la qualité des corpus comparables en termes de comparabilité pour améliorer la qualité des lexiques extraits → toujours d'actualité
 - Agir sur les méthodes d'extraction de lexiques bilingues à partir de corpus comparables → essoufflement même si il y a un renouveau avec les embeddings
 - Améliorer la qualité des résultats proposés par les techniques d'extraction de lexiques bilingues en tenant compte de leur contexte d'utilisation → peu étudié



MERCI