Traduction assistée par ordinateur et corpus comparables

Contributions à la traduction compositionnelle

Estelle Delpech

Laboratoire d'Informatique de Nantes Atlantique
équipe TALN
Directrice : Prof. Béatrice Daille
Co-encadrant : Prof. Emmanuel Morin

Soutenance de thèse
2 juillet 2013

Contexte : projet METRICC

Corpus comparables et :

- recherche d'information interlingue
- catégorisation multilingue
- aide à la traduction (LINGUA ET MACHINA, LINA)

Contexte : projet METRICC

Corpus comparables et :

- recherche d'information interlingue
- catégorisation multilingue
- **aide à la traduction (Lingua et Machina, LINA)**

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

VI. Conclusion générale

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
   Génération de traductions candidates
   Ordonnancement de traductions candidates

VI. Conclusion générale

Difficultés de la traduction technique [Darbelnet, 1979, Durieux, 2010]

Terminologie Notions du domaine, termes associés

- *chemotherapy* → *chimiothérapie*
- *neoangiogenesis* → *néoangiogénèse*

"Mise en discours"

- constructions syntaxiques, sous-catégorisation spécifiques
- usages stylistiques
- vocabulaire de soutien :
  - *patient-centred*→ *centré sur le patient*
  - *randomly* → *de manière randomisée*
- variation

Perspective de travail : acquisition de *lexiques* spécialisés bilingues

- Aspects non considérés : syntaxe, style
- Recherche d'équivalences traductionnelles :
  - Unités à traduire : toute une unité lexicale dont la traduction n'existe pas dans le dictionnaire généraliste
  - Perspective d'enrichissement
  - Prise en compte de la variation

Des corpus parallèles aux corpus comparables

- Historiquement : lexiques extraits de traductions passées (corpus parallèles)
- Limite : nouveaux domaines

Corpus comparables spécialisés

Ensemble de textes en langue L1 et L2 qui traitent d'une même thématique relative à un domaine de connaissance sans être en relation de traduction

Usage des corpus comparables en traduction technique

- Qualité reconnue par les experts de la traduction [Zanettin, 1998, Mc Enery and Xiao, 2007] :
- Usage “artisanal” et pédagogique
- Outils spécifiques existants
  - quelques prototypes universitaires [Bennison and Bowker, 2000, Sharoff et al., 2006]
  - pas d’outil commercial

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

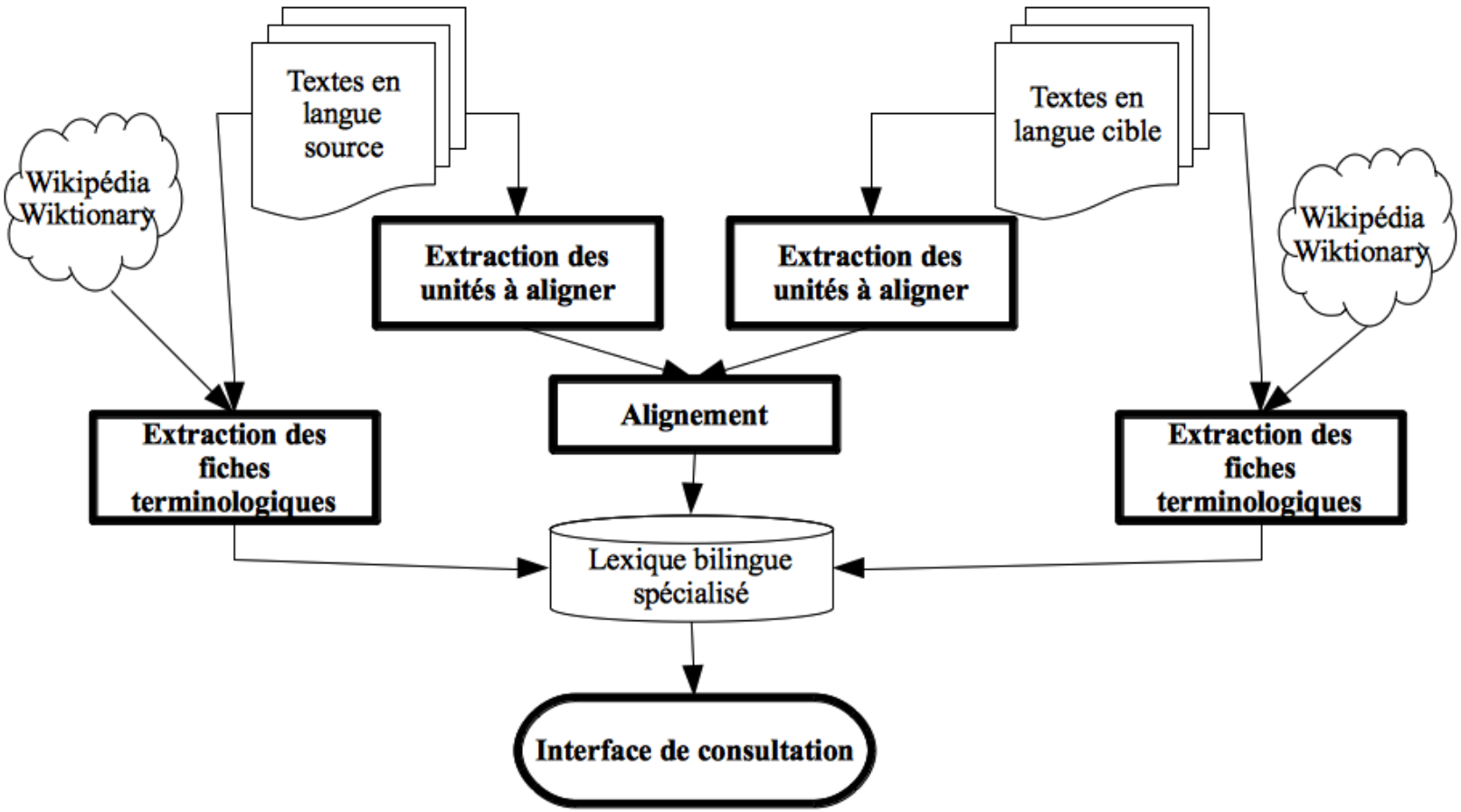
III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

VI. Conclusion générale

Architecture de l'extracteur

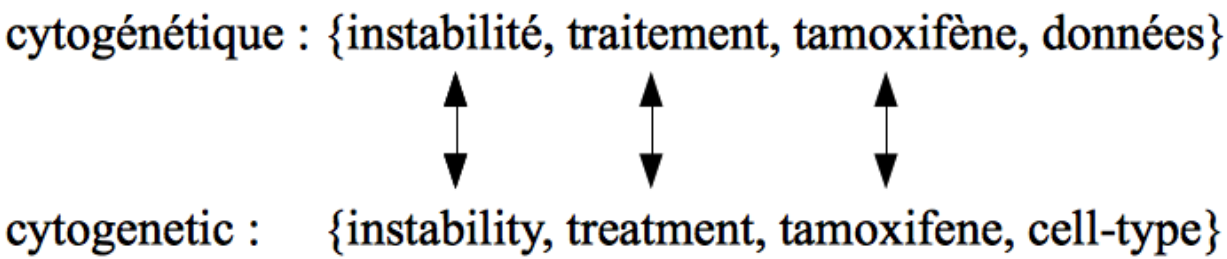


Extraction des unités à aligner

- Unités polylexicales : groupes nominaux et verbaux extraits par l'extracteur "terminologique" de Lingua et Machina
- Unités monolexicales (adjectif, verbe, nom, adverbe)

Méthode d'alignement

- Approche distributionnelle [Rapp, 1999, Fung, 1997] : deux mots de sens proche tendent à apparaître dans des contextes similaires



Implantation

- Diverses améliorations et variantes proposées [Déjean and Gaussier, 2002, Sadat et al., 2003, Morin et al., 2004, Prochasson, 2010, Hazem and Morin, 2012]
- Implantation basique avec adaptation aux unités polylexicales [Morin et al., 2004] et filtre sur les catégories grammaticales [Sadat et al., 2003]
- Résultats : 60% des unités à traduire avec une traduction correcte parmi les 20 premiers candidats

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

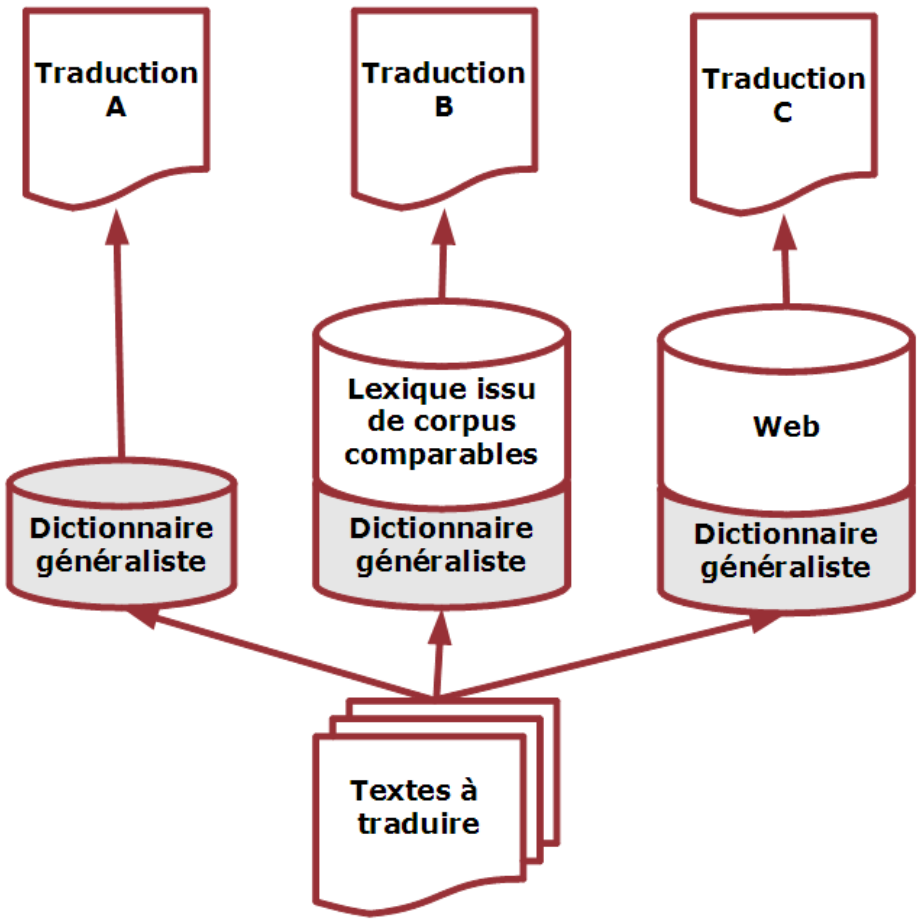
V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

VI. Conclusion générale

Méthodologie d'évaluation

- But : déterminer dans quelle mesure le lexique bilingue permet d'aider les traducteurs
- Méthode : comparaison de la qualité des traductions produites avec / sans les corpus comparables

Méthodologie d'évaluation



Méthodologie d'évaluation

- Objet évalué : expressions problématiques
- Mesure : % de traductions exactes, acceptables, fausses

Conditions expérimentales

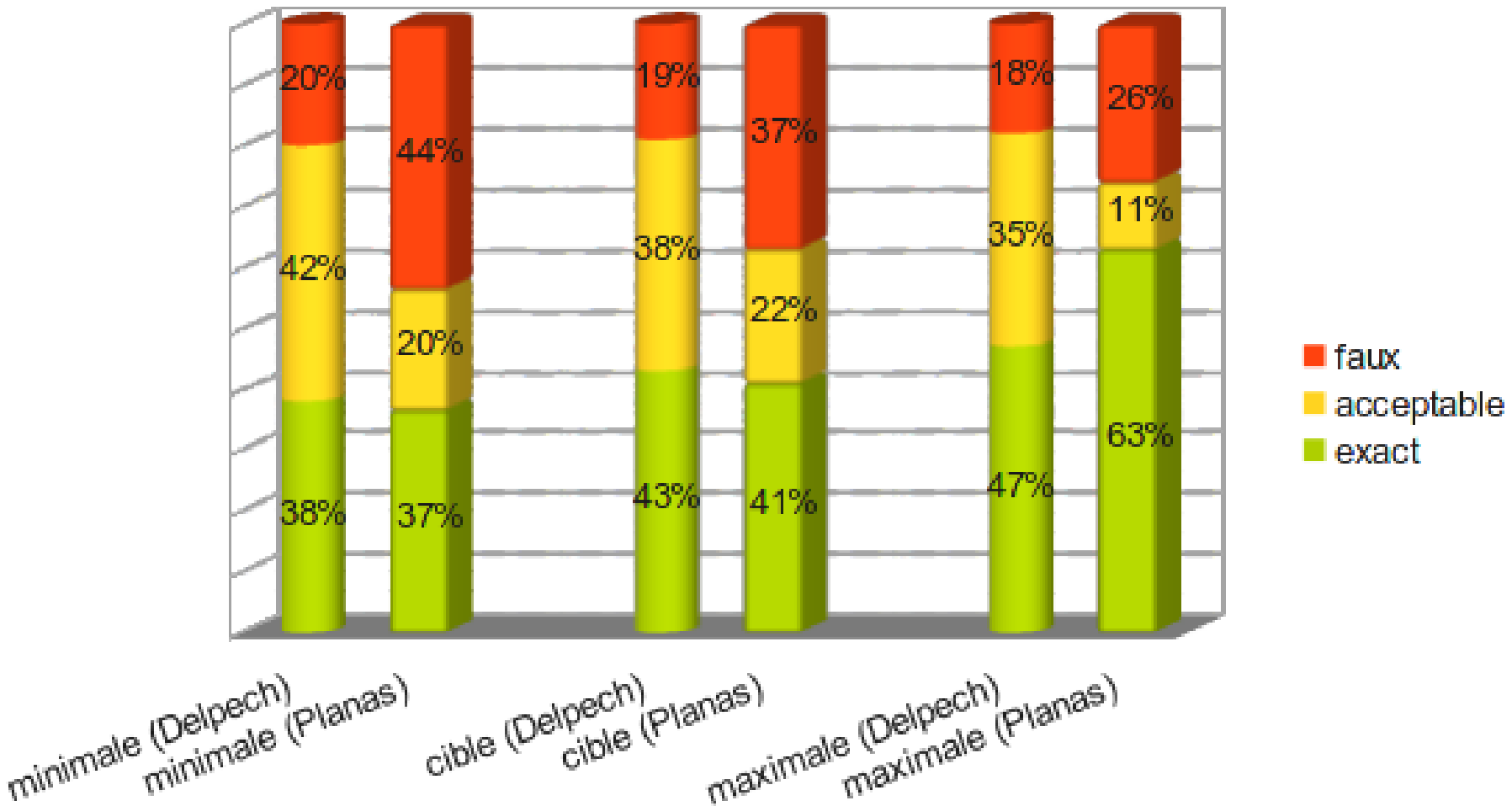
Première expérimentation visant à éprouver la méthode

- 3 traducteurs dont deux étudiant-e-s M2
- 2 thématiques : cancer du sein, sciences de l'eau
- Thématique sciences de l'eau trop vaste $\Rightarrow$pas exploitable

Évaluation finale [Planas, 2011]

- 20 étudiants-traducteurs de M1
- Données cancer du sein

Résultats


20%
42%
38%
44%
20%
37%
19%
38%
43%
37%
22%
41%
18%
35%
47%
26%
11%
63%
faux
acceptable
exact
minimale (Delpech)
minimale (Planas)
cible (Delpech)
cible (Planas)
maximale (Delpech)
maximale (Planas)

Difficultés d'usage

- Résistance au changement
  - ⇒ formation et recueil des besoins
- Trop de termes sources non couverts
  - ⇒ collecte du corpus
- Pas assez d'information pour choisir la bonne traduction
  - ⇒ contextualiser les traductions
- Trop de traductions candidates
  - ⇒ diminuer le nombre de traductions
  - ⇒ approche compositionnelle [Morin and Daille, 2010]

Difficultés d'usage

- Résistance au changement
  - ⇒ formation et recueil des besoins
- Trop de termes sources non couverts
  - ⇒ collecte du corpus
- Pas assez d'information pour choisir la bonne traduction
  - ⇒ contextualiser les traductions
- Trop de traductions candidates
  - ⇒ diminuer le nombre de traductions
  - ⇒ **approche compositionnelle** [Morin and Daille, 2010]

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

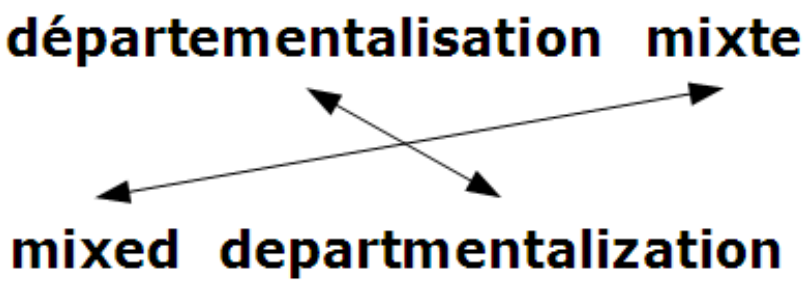
VI. Conclusion générale

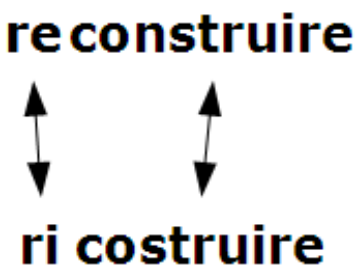
Principe de la traduction compositionnelle

Principe de compositionalité : "Le sens du tout est fonction du sens de ses constituants" [Keenan and Faltz, 1985, pp. 24-25].

Adaptation à la traduction : La **traduction** du tout est fontion de la **traduction** de ses constituants.

Exemples de traductions possibles





Difficultés

Divergence morpho-syntaxique :

anti-cancer → anti-cancéreux

Divergence lexicale :

traduction automatique → machine translation

Fertilité :

hysterectomy →ablation de l'utérus

Variation terminologique :

mixed departmentalization → départementalisation
mixte, structuration mixte

Aspects pas ou peu traités

- Fertilité
- Termes monolexicaux : approches spécifiques à un type de construction morphologique
  - $\text{prefixe}_1 + \text{base}_2 \rightarrow \text{préfixe}_1 + \text{base}_2$
- Ordonnancement / sélection des traductions : filtres simples ou pas adaptés

Propositions

- Termes monolexicaux : être moins spécifique sur les structures morphologiques
- Traiter la fertilité par l'alternance morphème libre / morphème lié
  - $cyto_1\, toxic_2 \rightarrow toxique_2\ (pour\ les)\ cellules_1$
- Explorer l'apport des critères d'ordonnancement et leur combinaison

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

VI. Conclusion générale

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

VI. Conclusion générale

Fonctionnement de base

$$
\begin{aligned}
\textit{Traduire}(\text{“ab”}) : & \\
&= \mathcal{S}(\mathcal{R}(\mathcal{T}(\mathcal{D}(\text{“ab”})))) \\
&= \mathcal{S}(\mathcal{R}(\mathcal{T}(\{\mathsf{a}, \mathsf{b}\}))) \\
&= \mathcal{S}(\mathcal{R}(\{\mathcal{T}(\mathsf{a}) \times \mathcal{T}(\mathsf{b})\})) \\
&= \mathcal{S}(\mathcal{R}(\{\mathrm{A}, \mathrm{B}\})) \\
&= \mathcal{S}(\{\mathrm{A}, \mathrm{B}\}, \{\mathrm{B}, \mathrm{A}\}) \\
&= \text{“BA”}
\end{aligned}
$$

Décomposition

- Peu de règles :
  - appariemment entrées ressources, contraintes longueur
- Tous les découpages possibles
  - $non\text{-}cytotoxic \rightarrow \{non,\ cyto,\ toxic\},\ \{noncyto,\ toxic\},\ \{non,\ cytotoxic\},\ \{noncytotoxic\}$

Traduction

- Équivalences traductionnelles entre morphèmes libres et liés
  - *cyto* → *cellule* : *cytotoxique* → *toxique pour les cellules*
- Nombreuses ressources : familles morphologiques, synonymes, cognats
  - *available* → *disponible* → *disponibilité* : *bioavailable* → *biodisponibilité*
  - *anastrozole-associated* → *associé à de l'anastrozole*
- Stratégie de repli
  - *confusingly* → *confusing* → *confondre*

Recomposition

- Permutation :
  - $pathophysiological \rightarrow physiopathologique$
- Tous les concaténations possibles :
  - $\{non,\ toxique,\ cellule\}$: $\{non,\ toxique,\ cellule\}$, $\{nontoxique,\ cellule\}$, $\{non,\ toxiquecellule\}$, $\{nontoxiquecellule\}$

Sélection

- Projection de patrons
  - $toxique\ .^*\ cellule \rightarrow toxique\ pour\ les\ cellules$

Données expérimentales

- Domaine cancer du sein, EN → FR, EN → DE
- ≃ 1800 unités monolexicales morphologiquement construites
  - aucune n'est traduisible avec le dictionnaire généraliste
- Ressources existantes : dictionnaire généraliste et synonymes
- Ressources manuelles : traductions morphèmes
- Ressources automatiques : familles morphologiques [Porter, 1980], cognats [Hauer and Kondrak, 2011]

Méthodologie d'évaluation

- Evaluation a priori : capacité de l'algorithme à reproduire un lexique existant, favorable, orienté terminologie
- Evaluation a posteriori : qualité des résultats obtenus en situation d'usage, orienté aide à la traduction
  - Annotation manuelle des sorties du système ;
  - EXACT, ACCEPTABLE, PROCHE, FAUX

Méthodologie d'évaluation

- Evaluation a priori : capacité de l'algorithme à reproduire un lexique existant, favorable, orienté terminologie
- **Evaluation a posteriori** : qualité des résultats obtenus en situation d'usage, orienté aide à la traduction
  - Annotation manuelle des sorties du système ;
  - EXACT, ACCEPTABLE, PROCHE, FAUX

Méthodologie d'évaluation

- Evaluation a priori : capacité de l'algorithme à reproduire un lexique existant, favorable, orienté terminologie
- **Evaluation a posteriori** : qualité des résultats obtenus en situation d'usage, orienté aide à la traduction
  - Annotation manuelle des sorties du système ;
  - **exact, acceptable**, PROCHE, FAUX

Mesures d'évaluation référence a priori

Couverture : capacité à générer une traduction candidate

Précision : capacité à proposer une traduction correcte parmi les traductions générée

Utilisabilité : capacité à générer une traduction candidate et correcte

Aspects évalués

- Généricité du modèle
- Ressources linguistiques
- Fertilité

Généricité du modèle : méthodes testées

- Préfixation : $pretreatment \rightarrow pré\text{-}traitement$
- Composition savante : $hypercalcaemia \rightarrow hypercalcémie$
- Composition populaire : $acute\text{-}phase \rightarrow Akutphase,\ akuten\ Phase$
- Cognat : $t\text{-}test \rightarrow t\text{-}Test$

Généricité du modèle : résultats

- Composition savante, préfixation : méthodes très précises (>0.92) mais petite couverture (<0.03)
- Composition populaire : petite couverture, moyennement précise (0.62 à 0.65)
- Cognats : meilleure couverture (0.10 à 0.13), précision moyenne à bonne (0.66 à 0.81)
- Notre méthode :
  - large couverture : 0.36 à 0.40
  - précision moyenne : 0.68 à 0.56
  - utilisabilité meilleure : 0.20 à 0.28 vs. cognats 0.07 à 0.10

Apport des traductions fertiles

- Traductions fertiles nettement moins précises (-0.20 à -0.39)
  - Combinaison aux traductions non fertiles intéressante (+6 à 10 points utilisabilité)

Bilan

- Fort gain en couverture, baisse limitée de la précison
  $\Rightarrow$ utilisabilité meilleure
- Limites :
  - fertilité sémantique : *snorkeling* $\rightarrow$ *plongée avec tuba*
- Perspectives :
  - compression : *après la ménopause* $\rightarrow$ *post-menopause*
  - termes polylexicaux : *cytogenetic instability* $\rightarrow$ *instabilité génétique des cellules*

Bilan

- Méthode bien adaptée aux corpus comparables spécialisés ...
  - peu d'a priori sur la structure du terme cible
  - variantes morphologiques
  - usage des cognats
- ... mais bruitée $\Rightarrow$ nécessite un filtrage

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

VI. Conclusion générale

Ordonnancement de traductions candidates

- Partie exploratoire
- Apports :
  - nouveaux critères
  - comparaison
  - combinaison (learning-to-rank)

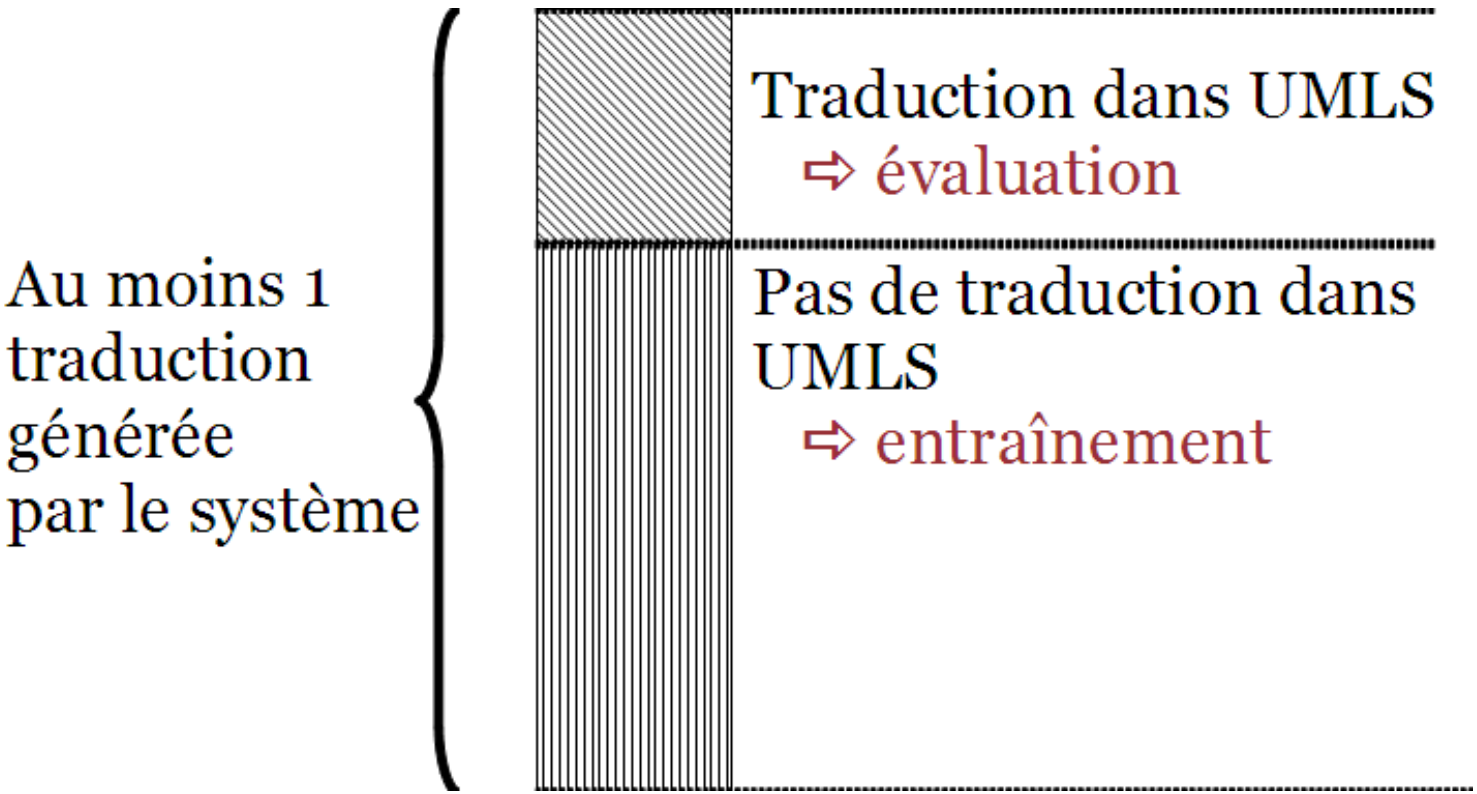
Critères

- F : fréquence traduction candidate
- C : similarité des contextes
- P : probabilité de traduction des partie du discours
- M : fiabilité des modes de traductions

Expériences

- Chaque critère pris isolément
- Combinaison non pondérée : $F+\ C+\ P+\ M$
- Combinaison pondérée : $\alpha F + \beta C + \gamma P + \delta M$
- Apprentissage modèles d'ordonnancement, famille *list-wise* :
  - AdaRank, Lambda MART : boosting
  - Coordinate Ascent : modèle linéaire

Données expérimentales



Méthode d'évaluation

- Précision sur le TopN : parmi les termes sources avec au moins 1 traduction candidate, % de ceux avec une traduction correcte parmi les N premières traductions
- Classement fonction de la précision sur le Top1 puis 2 puis 3

Résultats

- Comparaison des critères :
  - Contextes : moins bon critère (0.80 à 0.88 Top1)
  - Fiabilité des modes de traduction : meilleur critère (0.82 à 0.93 Top1)
- Meilleures méthodes (0.85 à 0.93, +5 à 9 points vs. aléatoire, Top1) :
  - Combinaison non pondérée
  - Combinaison pondérée
  - Coordinate Ascent, AdaRank

Bilan et perspectives

- Nécessité de montrer la significativité des résultats
- Globalement : combinaison intéressante, pas d'apport marqué des modèles de learning-to-rank (peu de critères)
- Autres critères : différence de fréquence, modèle de langue...
- Comment intégrer des données parallèles généralistes, d'autres domaines, d'autres langues ?
  - apprentissage à partir des traductions du dictionnaire généraliste et des cognats
  - poids valables pour tous les couples de langues...

Plan

I. Problématique : TAO et corpus comparables

II. Implantation d'un extracteur de lexiques bilingues à partir de corpus comparables

III. Evaluation applicative

IV. Approches compositionnelles

V. Traduction morpho-compositionnelle
- Génération de traductions candidates
- Ordonnancement de traductions candidates

VI. Conclusion générale

Bilan

- Développement d'un prototype d'extracteur de lexiques bilingues spécialisés à partir de corpus comparables [Delpech and Daille, 2010]
- Expérimentation de l'approche "classique" dans le cadre applicatif de la TAO [Delpech, 2011, Delpech, 2012] :
- Contributions à la traduction compositionnelle [Delpech et al., 2012b, Delpech et al., 2012a]:
- Communications et démonstrations logicielles [Delpech, 2010a, Delpech, 2010b, Brown de Colstoun et al., 2011]

Peut-on extraire des lexiques bilingues réellement **utilisables** par les traducteurs à partir de corpus comparables ?

- Approche compositionnelle :
  - réduit le nombre de traductions mais cantonnée aux éléments au sens compositionnel (60% d'après [Namer and Baud, 2007])
  - 20% à 28% avec une traduction correcte
- Approche distributionnelle :
  - forte couverture mais lexiques trop ambigus (60% sur le Top20, évaluation a priori)

Très ambitieux en l'état actuel

- Lexique obtenu :
  - une petite partie avec une traduction correcte sur le Top1 ou Top2
  - une partie avec traduction correcte sur Top 20
  - majeure partie sans traduction
- Difficile d'augmenter le corpus
  - thématique fine, forte comparabilité
  - peu de textes spécialisés
- Difficulté inhérente au corpus :
  - seulement une partie du vocabulaire en commun

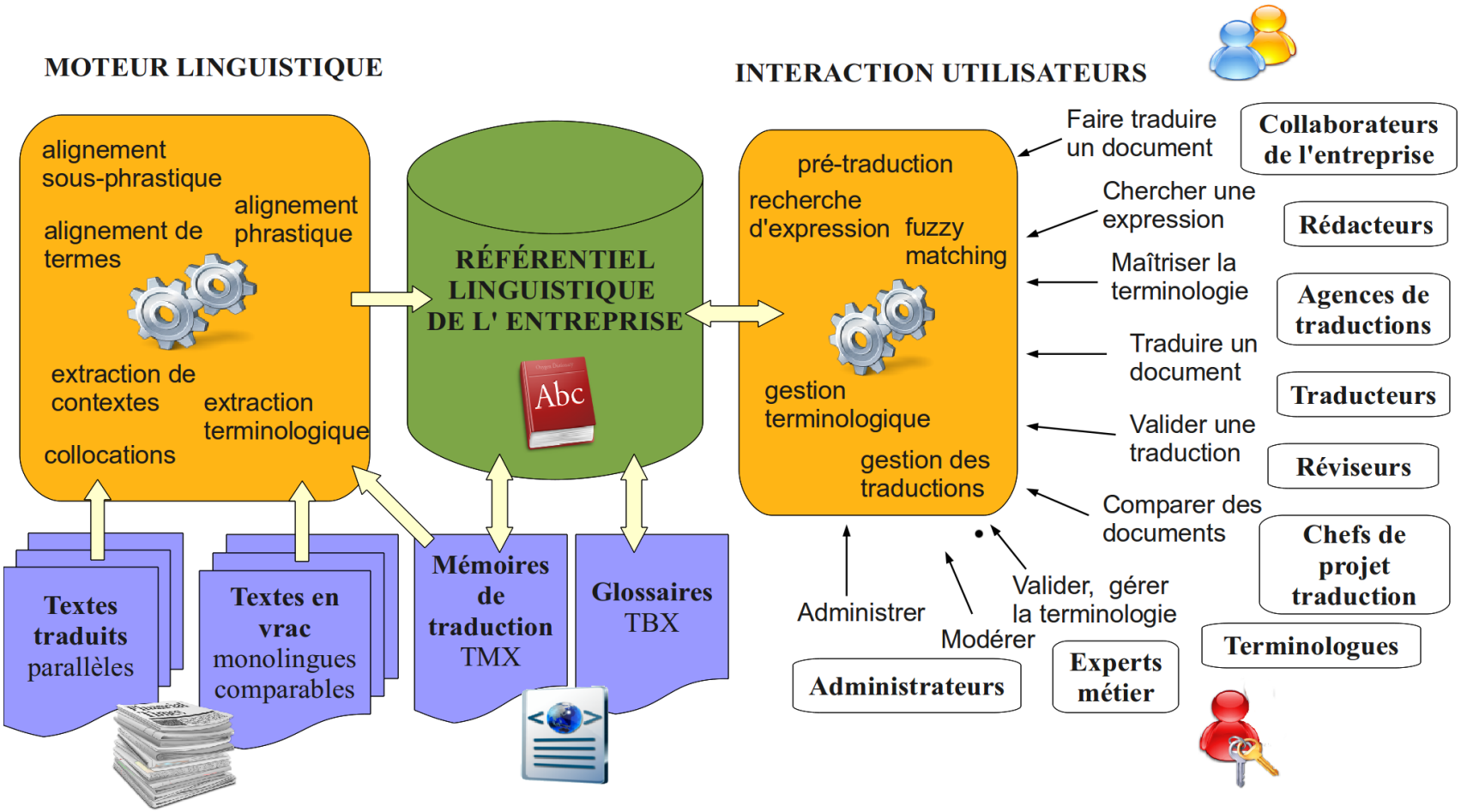
Perspectives

- Ne pas exagérement focaliser sur l'extraction d'alignements
- Aider à l'exploration de corpus comparables par de multiples manières :
  - extraction, alignement de contextes pertinents
  - outils de recherche avancés
  - travailler avec les traducteurs : automatiser les techniques

$\Rightarrow$ Projet CRISTAL : LINA, Lingua et Machina, CLLE-ERSS, Faculté de Traduction et d'Interprétation

Merci pour votre attention

Libellex


MOTEUR LINGUISTIQUE
alignement sous-phrastique
alignement de termes
alignement phrastique
extraction de contextes
extraction terminologique
collocations
RÉFÉRENTIEL LINGUISTIQUE DE L' ENTREPRISE
Abc
INTERACTION UTILISATEURS
pré-traduction
recherche d'expression
fuzzy matching
gestion terminologique
gestion des traductions
Faire traduire un document
Collaborateurs de l'entreprise
Chercher une expression
Rédacteurs
Maîtriser la terminologie
Agences de traductions
Traduire un document
Traducteurs
Valider une traduction
Réviseurs
Comparer des documents
Chefs de projet traduction
Valider, gérer la terminologie
Terminologues
Administrer
Modérer
Administrateurs
Experts métier
Textes traduits parallèles
Textes en vrac monolingues comparables
Mémoires de traduction TMX
Glossaires TBX

Méthode d'alignement

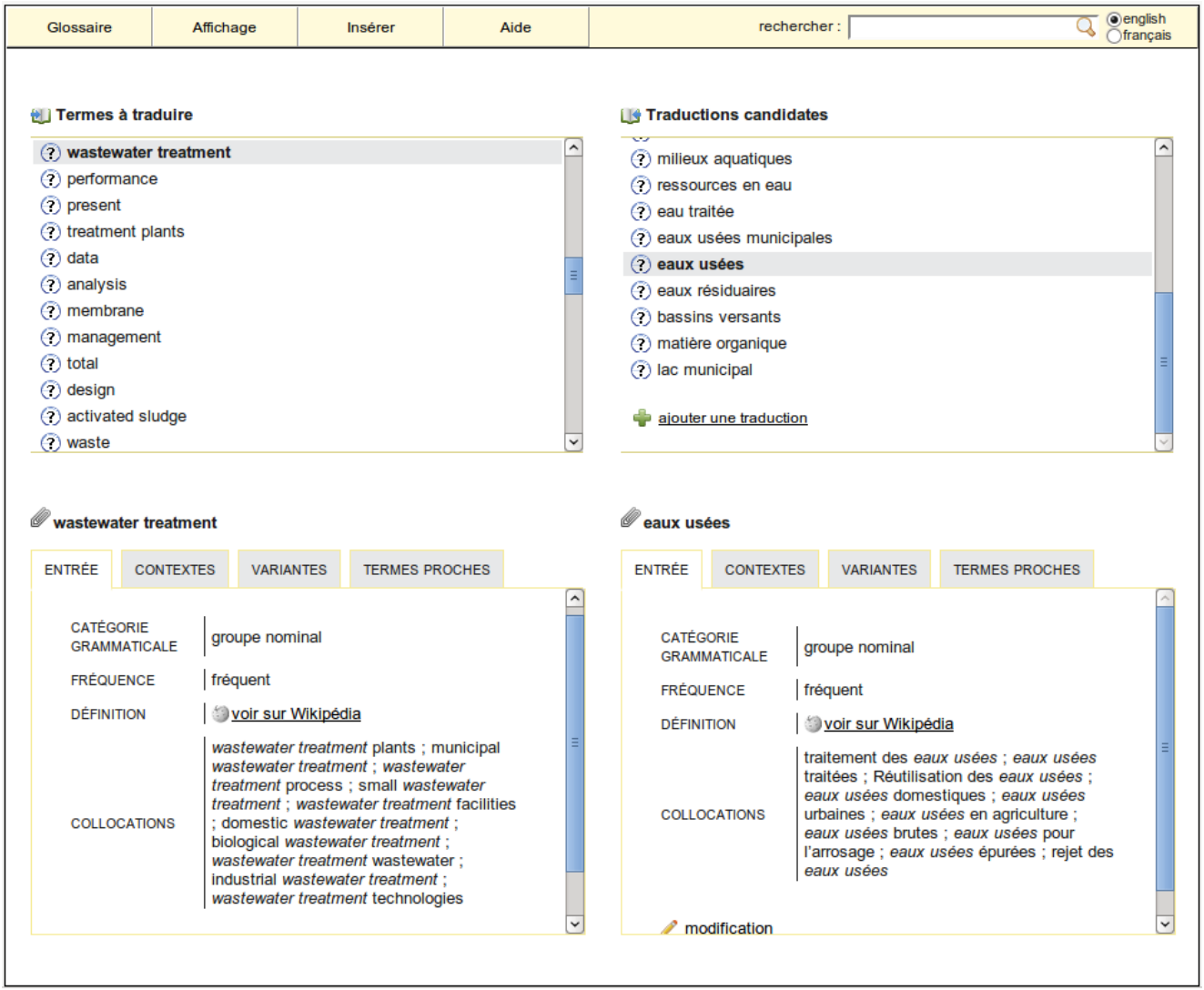
1. Collecte des contextes (vecteurs)
   $\overrightarrow{cytogénétique} = \{instabilité,\ traitement,\ tamoxifène...\}$
   $\overrightarrow{cytogenetics} : \{instability,\ treatment,\ tamoxifene...\}$

2. Traduction à l'aide d'un dictionnaire généraliste
   $\overrightarrow{cytogénétique} : \{instability,\ treatment,\ ?...\}$

3. Comparaison des contextes
   $similarité(\overrightarrow{cytogénétique}, \overrightarrow{cytogenetics}) = 0.75$

4. Sélection des candidats
   $cytogénétique \rightarrow cytogenetics\ (0.75),\ genetics\ (0.70)...$

Interface de consultation



Fertilité

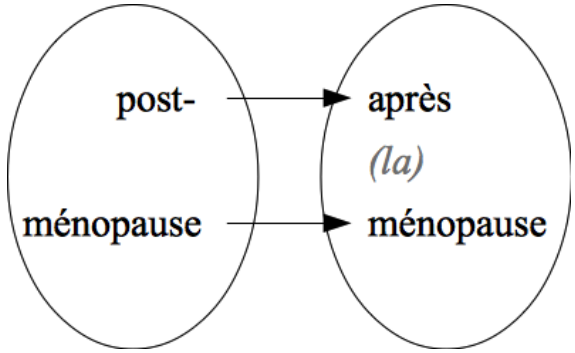
Traduction fertile *Soit deux ensembles disjoints S et C où S est un ensemble de termes sources et C est un ensemble de termes cibles. Soit la relation de traduction $T \subseteq S \times C$ et la fonction $l(x)$ indiquant le nombre de mots lexicaux du terme x. L'ensemble des traductions fertiles F est défini comme $\{(s, c)|(s, c) \in T \text{ et } l(c) > l(s)\}$.*

Exemples :

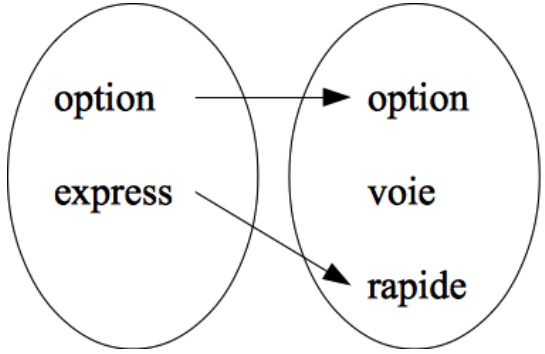
- *post-menopause* → *après (la) ménopause*
- *option express* → *option voie rapide*
- *snorkeling* → *plongée (avec) tuba*

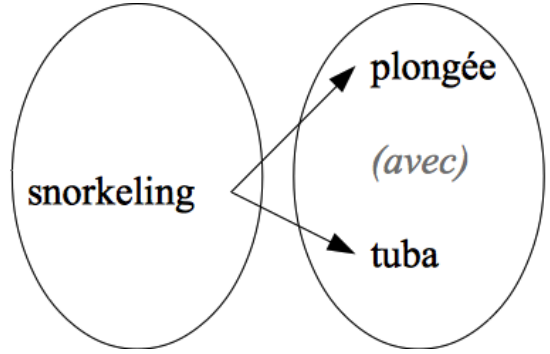
Fertilité de surface et fertilité sémantique

Surface



Sémantique





Généricité I

	C	P_E	U_E	P_{EA}	U_{EA}
Composition savante (18%)	,03	**,95**	,03	**1**	,03
Cognat	,13	,66	,08	,81	,10
Composition populaire (48 %)	,05	,63	,03	,65	,03
Préfixation (31%)	,02	,90	,02	,97	,02
Notre méthode	**,40**	,59	**,24**	,69	**,28**

Table: anglais → français

Généricité II

	C	P_E	U_E	P_{EA}	U_{EA}
Composition savante (18%)	,03	**,96**	,02	**,98**	,02
Cognat	,10	,58	,06	,66	,07
Composition populaire (49 %)	,04	,55	,02	,62	,03
Préfixation (32%)	,03	,86	,02	,92	,03
Notre méthode	**,36**	,48	**,17**	,56	**,20**

Table: anglais → allemand

Généricité du modèle : discussion

- Variation morphologique : *pretreatment* → *prétraiter*, *cardiotoxicity*→ *cardiotoxique*, *time-consuming* → *consommateur de temps*
- Fertilité : *pretreatment* → *avant le traitement*, *hypercalcaemia* → *zu viel calcium in das blut*
- Cognats : *aromatase-inhibiting* → *hemmung der aromatase*'*inhibition de l'aromatase*'
- Suffixes : *colorless* → *sans colorant*, *randomly* → *(de) manière randomisée*
- Stratégie de repli : *ribosome* → *ribosomique*

Ressources linguistiques : comparaisons effectuées

- Base : dictionnaire généraliste et table de traduction des morphèmes
- Base + familles morphologiques
- Base + synonymes
- Base + cognats
- Toutes les ressources

Apport des ressources linguistiques

Système de base : dictionnaire généraliste et table de traduction des morphèmes

- Synonymes : pas adaptés ($bloodstream \rightarrow courant\ sanguin \rightarrow circulation\ sanguine$)
- Familles morphologiques : +0.09 à 0.11 de couverture ; +0.04 à 0.06 d'utilisabilité
- Cognats : + 0.12 de couverture ; +0.06 à 0.09 d'utilisabilité
- Combinaison : + 0.17 à 0.24 couverture ; +0.10 à 0.16 utilisabilité

Ressources linguistiques I

	C	P_E	U_E	P_{EA}	U_{EA}
Base	,16	**,73**	,12	**,77**	,12
Base + dictionnaire de cognats	,28	,71	,19	**,77**	,21
Base + familles morphologiques	,27	,56	,15	,66	,18
Base + dictionnaire synonymes	,17	,69	,12	,72	,13
Toutes les ressources	**,40**	,59	**,24**	,69	**,28**

Table: anglais → français

Ressources linguistiques II

	C	P_E	U_E	P_{EA}	U_{EA}
Base	,15	**,60**	,09	**,63**	,10
Base + dictionnaire de cognats	,27	,56	,15	,61	,16
Base + familles morphologiques	,24	,48	,12	,57	,14
Base + dictionnaire synonymes	,17	,55	,09	,60	,10
Toutes les ressources	**,36**	,48	**,17**	,56	**,20**

Table: anglais → allemand

Traductions fertiles I

	C	P_E	U_E	P_{EA}	U_{EA}
Traductions non fertiles	,24	**,58**	**,14**	**,75**	**,18**
Traductions fertiles	,24	,52	,12	,55	,13

Traductions non fertiles	,24	,58	,14	**,75**	,18
Toutes les traductions	**,40**	**,59**	, **24**	,69	**,28**

Table: anglais → français

Traductions fertiles II

	C	P_E	U_E	P_{EA}	U_{EA}
Traductions non fertiles	**,24**	**,58**	**,14**	**,69**	**,16**
Traductions fertiles	,20	,26	,05	,30	,06

Traductions non fertiles	,24	**,58**	,14	**,69**	,16
Toutes les traductions	**,36**	,48	**,17**	,56	**,20**

Table: anglais → allemand

Résultats anglais → français

	Top1	Top2	Top3	RPM
Meilleure précision possible	,94	,94	,94	1
Combinaison non pondérée	**,928**	**,94**	**,94**	**2**
Combinaison pondérée	**,928**	**,94**	**,94**	**2**
Coordinate Ascent	**,928**	**,94**	**,94**	**2**
Lambda MART	**,928**	**,94**	**,94**	**2**
M	**,928**	**,94**	**,94**	**2**
F	,916	,928	**,94**	3
AdaRank	,892	,904	,928	4
P	,892	,904	,928	4
C	,88	,904	,928	4
Aléatoire	,836	,898	,928	13

Résultats anglais → allemand

	Top1	Top2	Top3	RPM
Meilleure précision possible	,879	,879	,879	1
Combinaison pondérée	**,848**	**,879**	**,879**	**2**
Lambda MART	**,848**	,864	,864	**5**
Combinaison non pondérée	,833	,864	**,879**	**3**
Coordinate Ascent	,833	,864	**,879**	**3**
F	,833	,848	**,879**	**3**
AdaRank	,833	,848	,848	17
P	,833	,848	,848	17
M	,818	,864	**,879**	**3**
C	,803	,864	,864	28
Aléatoire	,77	,832	,846	28

Comparaison avec [Claveau and Kijak, 2011]

	P_E	R_E	$F1_E$	# exemples	table morphèmes
Delpech EN-FR	,93	,61	,74	1 970	242 → 1001
Delpech EN-DE	,85	,62	,72	1 829	250 → 1081
Claveau FR-JA	,89	,64	,74	6 400	0

Références I

Bennison, P. and Bowker, L. (2000).

Designing a tool for exploiting bilingual comparable corpora.
In *Proceedings of LREC 2000*, Athens, Greece.

Brown de Colstoun, F., Delpech, E., and Monneret, E. (2011).

Libellex : une plateforme multiservices pour la gestion des contenus multilingues.
In Lafourcade, M. and Prince, V., editors, *Actes de la 18ème conférences sur le traitement automatique des langues naturelles*, volume 2, page 319, Montpellier, France.

Claveau, V. and Kijak, E. (2011).

Morphological analysis of biomedical terminology with Analogy-Based alignment.
In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 347–354, Hissar, Bulgaria.

Darbelnet, J. (1979).

Réflexions sur le discours juridique.
Meta : journal des traducteurs / Meta: Translator's Journal, 24(1):26–34.

Déjean, E. and Gaussier, E. (2002).

Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables.
Lexicometrica, Alignement lexical dans les corpus multilingues, pages 1–22.

Delpech, E. (2010a).

Bilingual terminology mining.
In *The 4th Intensive Summer school and collaborative workshop on Natural Language Processing (Franco-Thai Workshop 2010)*, Bangkok, Thaïlande.

Delpech, E. (2010b).

Libellex, environnement de gestion collaborative en ligne de terminologie au sein de communautés fermées.
In *Terminologie & Ontologie : Théories et applications (TOTh)*, Annecy, France.

Références II

Delpech, E. (2011).

Evaluation of terminologies acquired from comparable corpora : an application perspective.
In Pedersen B.S., Nešpore G., S. I., editor, *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, volume 11 of *NEALT Proceedings Series*,, pages 66–73, Riga, Latvia.

Delpech, E. (2012).

Un protocole d'évaluation applicative des terminologies bilingues destinées à la traduction spécialisée.
Revue des Nouvelles Technologies de l'Information (RNTI) - Numéro spécial : Evaluation des méthodes d'Extraction de Connaissances dans les Données (Eval'ECD).

Delpech, E. and Daille, B. (2010).

Dealing with lexicon acquired from comparable corpora : validation and exchange.
In *Proceedings of the 2010 Terminology and Knowledge Engineering Conference (TKE 2010)*, pages 211–223, Dublin, Ireland.

Delpech, E., Daille, B., Morin, E., and Lemaire, C. (2012a).

Extraction of domain-specific bilingual lexicon from comparable corpora: a compositional translation and ranking.
In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 745–762, Mumbai, Inde.

Delpech, E., Daille, B., Morin, E., and Lemaire, C. (2012b).

Identification of fertile translations in medical comparable corpora: a morpho-compositional approach.
In *Proceedings of the 10th biennial conference of the Association for Machine Translation in the Americas*, San Diego, California.

Durieux, C. (2010).

Fondement didactique de la traduction technique.
La maison du dictionnaire, Paris, France.

Références III

Fung, P. (1997).

Finding terminology translations from non-parallel corpora.
In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.

Hauer, B. and Kondrak, G. (2011).

Clustering semantically equivalent words into cognate sets in multilingual lists.
In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 865–873, Chiang Mai, Thailand.

Hazem, A. and Morin, E. (2012).

ICA for bilingual lexicon extraction from comparable corpora.
In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, Istanbul, Turkey.

Keenan, E. L. and Faltz, L. M. (1985).

Boolean semantics for natural language.
Dordrecht, Holland.

Mc Enery, A. M. and Xiao, R. Z. (2007).

Parallel and comparable corpora: What is happening?
In G. Anderman, M. R., editor, *Incorporating Corpora: The Linguist and the Translator.*, Translating Europe, pages 18–31. Multilingual Matters, Clevedon, UK.

Morin, E. and Daille, B. (2010).

Compositionality and lexical alignment of multi-word terms.
In Rayson, P., Piao, S., Sharoff, S., Evert, S., and B., V., editors, *Language Resources and Evaluation (LRE)*, volume 44 of *Multiword expression: hard going or plain sailing*, pages 79–95. Springer Netherlands.

Morin, E., Dufour-Kowalski, S., and Daille, B. (2004).

Extraction de terminologies bilingues à partir de corpus comparables.
In *Actes de la 11ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 309–318, Fès, Maroc.

Références IV

Namer, F. and Baud, R. (2007).

Defining and relating biomedical terms: Towards a cross-language morphosemantics-based system.
International Journal of Medical Informatics, 76(2-3):226–33.

Planas, E. (2011).

Metricc : Rapport final sur l'évaluation de l'apport des lexiques bilingues pour la traduction.
Délivrable ANR n°28 lot 4.3, Université de Nantes, Nantes.

Porter, M. F. (1980).

An algorithm for suffix stripping.
Program, 14(3):130–137.

Prochasson, E. (2010).

Alignement multilingue en corpus comparables spécialisés : Caractérisation terminologique multilingue.
Thèse en informatique, Université de Nantes, Nantes.

Rapp, R. (1999).

Automatic Identification of Word Translations from Unrelated English and German Corpora.
In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

Sadat, F., Yoshikawa, M., and Uemura, S. (2003).

Learning bilingual translations from comparable corpora to Cross-Language information retrieval: Hybrid statistics-based and linguistics-based approach.
volume 11, pages 57–64, Sappro, Japan.

Sharoff, S., Babych, B., Rayson, P., Mudraya, P., and Piao, S. (2006).

ASSIST: automated semantic assistance for translators.
In *Proceedings to the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–142, Trento, Italie.

Références V

Zanettin, F. (1998).
Bilingual comparable corpora and the training of translators.
Meta : journal des traducteurs / Meta: Translator's Journal, 43(4):616–630.