Contextes riches en connaissances pour l'aide à la traduction - le projet CRISTAL

Cécile Fabre

SL02353X

Thématiques actuelles de la recherche en TAL

4/11/2013

Plan

1. Présentation du projet CRISTAL
   - Organisation et objectifs
   - La TAO
   - La notion de contexte
   - Les corpus comparables
2. Contextualiser pour le traitement automatique
3. Contextualiser pour le traducteur
4. Quels bons contextes pour le traducteur
   - Le bon exemple en lexicographie
   - Le contexte riche en connaissances en terminologie
   - Evaluer les besoins des traducteurs

Le projet CRISTAL

Contextes RIches en connaissanceS pour la TrAduction terminoLogique

- ANR (appel Contenus et Interactions) 2012-2015
- Partenaires :
  - LINA, Nantes
    - Emmanuel Morin, Béatrice Daille, Emmanuel Planas
  - CLLE-ERSS
    - Anne Condamines, Cécile Fabre, Amélie Josselin-Leray, Josette Rebeyrolle / thèse Luce Lefeuvre
  - Faculté deTraduction et d'Interprétation (FTI), Genève
    - Aurélie Picton
  - Lingua et Machina, société spécialisée en TAO et gestion terminologique
    - François Brown de Colstoun

Objectifs du projet

- Dans le domaine de la traduction assistée, faciliter le travail du traducteur et permettre la production de dictionnaires terminologiques plus riches
- En particulier : extraire les contextes d'emploi des termes (en langue source et en langue cible) pour en faciliter :
  - la compréhension (appréhender leur sens exact)
  - la production (les employer correctement)
- Extraire des contextes riches pour le traducteur

La notion de contexte riche

- Deux types de contextes distingués au démarrage du projet :
  - Contextes orientés compréhension, **CRC conceptuels**
    - ***Pumice*** *is the term applied to vesicular ejecta*
    - *toutes sortes de roches altérées : ignimbrite, tuf,* ***ponce*** *et obsidienne*
    - *En revanche, son cratère sommital (ou* ***caldera****) pourtant impressionnant*

    => Marqueurs de relations conceptuelles entre termes
  - Contextes orientés usage, **CRC linguistiques**
    - *Fresh volcanic material and floating rafts of* ***pumice*** *also indicate submarine eruptions.*
    - *L'île fut alors recouverte de* ***ponce*** *et de cendres à plusieurs reprises.*
    - *Le Trou au Natron est une grande* ***caldera*** *d'environ 7km de diamètre*

    => Collocations, dimension phraséologique

Rôles des partenaires

- CLLE-ERSS et FTI :
  - Définir la notion de contexte riche pour les traducteurs
  - Tester leur validité auprès de ces utilisateurs
  - Proposer des patrons pour le repérage de contextes riches
- LINA :
  - Implémenter les procédures de sélection de contextes à partir des patrons fournis
- Lingua et Machina :
  - Intégrer ces procédures nouvelles dans la chaîne de traitement existante

La TAO

- Logiciels d’aide à la traduction spécialisée
- Tâche primordiale : acquérir la terminologie du domaine
- Manque de ressources bilingues terminologiques

Les pratiques des traducteurs

(enquête rapportée dans Blancafort et al. 2011)

- 74% utilisent des outils de TAO et de TA
  - TAO : Trados, Similis
  - TA commerciale : Systran, Language Weaver
  - TA en ligne : Google translate
- Mais 50% seulement constituent des corpus spécialisés
  - Leur utilisation est principalement manuelle :
    - 30% utilisent des concordanciers
    - 10% des outils de TAL
- 94% utilisent Google tout en le jugeant inadapté

Plateforme de TAO

- Aide à la traduction :
  - Identification de segments déjà traduits
  - Correspondance exacte ou approximative / *exact or fuzzy matching*
- Utilisation et aide à la constitution de mémoires de traduction
- Aide à la constitution de glossaires et terminologies bilingues

Plateforme Libellex

- Aide à la traduction : *fuzzy matching*

Français !</strong-1><strong-2> </strong-3><.>
<strong-0>Malgré le contexte économique actuel, les Banques Alimentaires espèrent que la générosité sans faille, dont les français font preuve depuis plus de 20 ans, sera cette année encore à la hauteur de leurs besoins.</strong-1><.>

cette année	this year	100
la générosité	generosity	100
plus de 20 ans	than 20 years	100
les Banques Alimentaires	the Food Banks	93
faille	fail	87

<.>
Document sans nom<.>
<img-0> <strong-1>Les 28 & 29 novembre 2008 : Collecte Nationale
</strong-2><.>
<strong-0>Les Banques Alimentaires comptent sur la grande générosité des Français !</strong-1><strong-2> </strong-3><.>
<strong-0>Malgré le contexte économique actuel, les Banques Alimentaires espèrent que la générosité sans faille, dont les français font preuve depuis plus de 20 ans, sera cette année encore à la hauteur de leurs besoins.</strong-1><.>
<strong-0>Les 90 000 bénévoles mobilisés pendant ces deux jours font le pari

Plateforme Libellex

- Aide à la constitution de ressources bilingues
- Proposition de traductions candidates

adjuvant (en)fr	1. substitutif
	2. adjuvant
	3. conservateur
	4. séquelle
	5. hormonal
	6. arsenal
	7. néoadjuver
	8. locorégiona
	9. initial
	10. systémique
	11. bénéfier

Traduction Assistée à partir de corpus comparables

- Passer des corpus alignés aux corpus comparables

"Comparable corpora consist of original texts in each language, matched as far as possible in terms of text type, subject matter and communicative function" (Altenberger et Berger 2002)

- Limites des corpus parallèles :
  - Bons résultats mais ressources rares
  - Données artificielles, risques de *translationese*
- Difficultés des corpus comparables :
  - Pas d'élément d'ancrage
  - Faible comparabilité

3 enjeux

- En amont : constituer un corpus de qualité pour améliorer la comparabilité
  - Critères qualitatifs :
    - Homogénéité : domaine, thème, type de discours
  - Critères quantitatifs :
    - Proportion des traductions possibles des termes du corpus source (resp. cible) qui se retrouvent dans le corpus cible (resp. source). (Li et Gaussier 2010)
- Pendant la phase d'alignement :
  - Pas d'alignement direct => comparabilité des contextes
  - Hypothèse : un mot et sa traduction ont tendance à apparaître dans les mêmes environnements lexicaux
- En aval :
  - Proposer des traductions candidates
  - Assister le choix du traducteur en lui fournissant des indices contextuels

Double rôle de la contextualisation dans le projet

- Pour le traitement automatique :
  - Réordonnancement automatique des traductions candidates

    => Si le terme à traduire et une des traductions candidates sont trouvés dans des contextes comparables, ils ont de fortes chances d'être liés

- Pour le traducteur :
  - Proposition de bons contextes

    => Les traductions candidates sont replacées dans des contextes suffisamment informatifs pour éclairer le choix du traducteur

Corpus d'étude

- Deux corpus comparables permettant de travailler sur la variation en genre, langue et domaine

	Vulgarisation	Articles scientifiques
Volcanologie	200 000 mots / langue	200 000 mots / langue
Cancer du sein	400 000 mots/langue	400 000 mots/langue

Plan

1. Présentation du projet CRISTAL
   - Organisation et objectifs
   - La TAO
   - La notion de contexte
   - Utilisation de corpus comparables en TAO
2. Contextualiser pour le traitement automatique
3. Contextualiser pour le traducteur
4. Quels bons contextes pour le traducteur
   - Le bon exemple en lexicographie
   - Le contexte riche en connaissances en terminologie
   - Evaluer les besoins des traducteurs

Extraire des traductions d'un corpus comparable

(Delpech et al. 2012)

- Traduction de termes complexes
- Principe = traduction compositionnelle :

  Les traductions candidates des termes complexes sont produites à partir de :

  - la traduction de leurs composants atomiques
  - la génération de toutes les combinaisons possibles

Delpech, E., Daille, B., Morin, E. et Lemaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking. Proceedings, *24th International Conference on Computational Linguistics (COLING 2012)*, pp. 745-761, Mumbai, India. 2012.

Traduction compositionnelle

A partir d'un corpus comparable :

1) Extraire les termes complexes de chaque corpus monolingue
   - Utilisation d'Acabit (Daille)

2) Utiliser un dictionnaire bilingue :
   - Traduction directe du mot complexe
   - Traduction compositionnelle :
     - Extraire la traduction des mots simples
     - Opérer toutes les combinaisons

3) Garder les traductions qui correspondent à des termes extraits ou qui sont dans le corpus cible

Traduction compositionnelle

- 3 sources d’échec :
  - Le terme n’est pas compositionnel
  - Un des composants n’est pas dans le dictionnaire
  - Aucune combinaison n’existe dans le corpus
- Solution : approche compositionnelle enrichie par des infos contextuelles
  - Le mot absent du dictionnaire est remplacé par son vecteur de contexte

  (Morin et Daille 2012) : context-based projection

Réordonnancement

- Production de plusieurs traductions candidates
- Plusieurs stratégies d'ordonnancement possibles :
  - (Delpech et al. 2012) utilisent 4 critères :
    - La fréquence de la traduction
    - La probabilité de la traduction en termes de POS
    - La similarité des contextes :
      - Vecteur du contexte de la traduction V1
      - Vecteur du contexte de la traduction V2
      - Traduction du vecteur de contextes V1 => V1'
      - Comparaison V1' / V2 (mesure de jaccard pondérée)
    - La validité de la source de traduction :
      - Dictionnaire général, spécialisé, *cognate matching*, variation lexicale, morphologique …

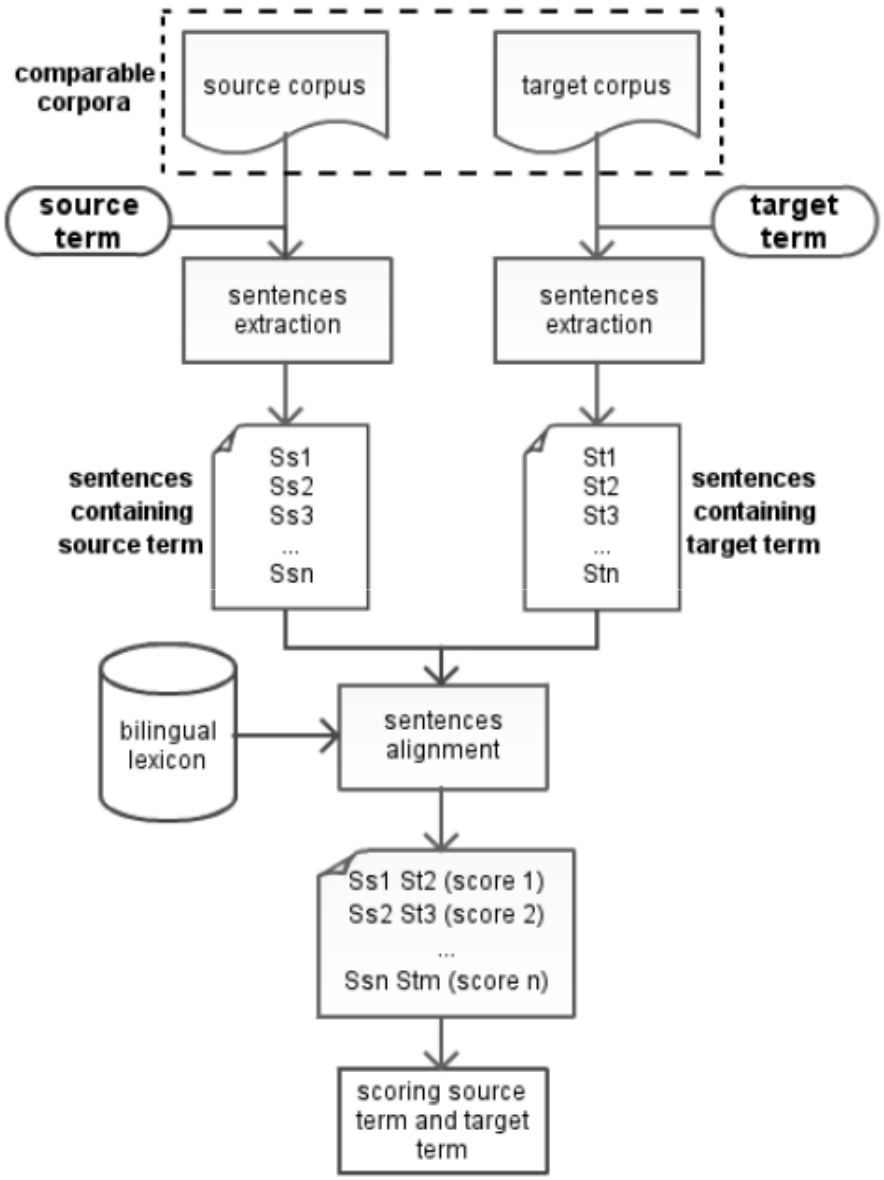
Réordonnancement

- Nouvelle méthode : recherche des meilleurs contextes comparables

(Harastani et al. 2013)

- Principes :
  - Le terme et sa traduction doivent apparaître dans des contextes comparables
  - Ces contextes comparables doivent être de « bons » contextes pour chacun des termes

Harastani, R.; Daille, B. & Morin, E. (2013), Ranking Translation Candidates Acquired from Comparable Corpora, *in 'Proceedings of the Sixth International Joint Conference on Natural Language Processing', Asian Federation of Natural Language Processing, Nagoya, Japan, pp. 401--409.*



(Harastani et al. 2013)

Réordonnancement des traductions candidates

- Méthode :
  - Extraction des meilleures phrases pour le terme source
  - Extraction des meilleures phrases pour toutes les traductions candidates
  - Recherche pour chaque phrase du terme source la phrase la plus comparable pour le terme cible

Phrases comparables

- = phrases qui partagent des données parallèles (recouvrement de mots, longues séquences équivalentes, noms composés)

Source sentence:
L'examen radiologique doit être associé à un examen clinique médical simultané, capable de détecter des tumeurs de très petites dimensions.

Target sentence:
There was no association between the tumor size detected during clinical examination mammography, MRI or histopathological analyses and presence of risidual disease.

Connected words:
(examen, examination), (clinique, clinical), (détecter, detected), (tumeurs, tumor), (dimensions, size)

(Harastani et al. 2013)

Qu’est-ce qu’un bon contexte pour un terme ?

- Ex : *tumor*

Un bon contexte (A)	Un mauvais contexte (B)
Chemotherapy was also administered to patients with smaller primary tumors with histological grade 2 or 3 or with negative hormone receptors.	The size of any captured image corresponding to the tumor was estimated.

Qu’est-ce qu’un bon contexte pour un terme ?

- Ex : *tumor*

Un bon contexte (A)	Un mauvais contexte (B)
Chemotherapy was also administered to patients with smaller primary tumors with histological grade 2 or 3 or with negative hormone receptors.	The size of any captured image corresponding to the tumor was estimated.

- (A) contient des mots qui sont très spécifiques au domaine considéré (cancer du sein).

Qu'est-ce qu'un bon contexte dans ce cas ?

- 2 critères utilisés pour estimer la qualité d'un contexte :
  - **Score de spécificité** d'un mot dans le domaine $ds(w)$
    - Fréquence relative dans le domaine / fréquence relative dans corpus général
  - **Score d'association** d'un mot avec *t* $assoc(w_i,t)$
    - Fenêtre de 7 mots autour de *t*
    - Vecteur des 30 mots de contexte les plus fortement associés à *t* (loglikelihood association) v_m
- Score d'une phrase S contenant *t* et les mots w_i, w_2, … w_n :

$$Score(S)=\sum_n^1 ds(w_i) + assoc_{(if\ wi \in v_m)}(w_i,t)$$

N.B. on ne conserve préalablement que les phrases contenant plus de 5 mots pleins

Dernière étape : alignement des phrases

- Alignement des *n* meilleures phrases de t avec les *n* meilleures phrases de chacune de ses traductions possibles
- Critères de comparabilité :
  - Similarité des 2 vecteurs de phrase par la mesure cosinus
  - Position des mots connectés par rapport à *t* et à *t'*
    - Plus les mots sont proches et plus les positions sont comparables, mieux c'est
  - Segment continu le plus long
  - Nombre de bigrammes communs
- Evaluation (corpus cancer français-anglais) :
  - P1 : 26,22% → 31,96%
  - P5 : 45,08% → 52,45%
  - P10 : 53,27% → 57,37%

Ex : *dépistage / screening*

Score de la phrase : 0.389

(@card@) le mammographie de **dépistage** effectuer dans le cadre du dépistage organiser du cancer du

genetics service , with at least one relative diagnose with breast cancer at less-than-or-equals , slant30 year also undergo **screening** for mutation in brca1/ 2

<connectedWords>

<sourceWord position="8">effectuer</sourceWord>
<targetWord position="19">undergo</targetWord>

<sourceWord position="16">cancer</sourceWord>
<targetWord position="12">cancer</targetWord>

</connectedWords>

Score de la phrase : 0.222

an retenir pour le **dépistage** organiser du cancer du sein , dépistage reposer sur un mammographie tout le deux an

the remain 9 woman (@card@ %) have their cancer detect within the breast **screening** service

<connectedWords>

<sourceWordposition="10 ">sein</sourceWord>
<targetWord position="15"> breast</targetWord>

<sourceWord position="8">cancer</sourceWord>
<targetWord position="11">cancer</targetWord>

</connectedWords>

Plan

1. Présentation du projet CRISTAL
   - Organisation et objectifs
   - La TAO
   - La notion de contexte
   - Utilisation de corpus comparables en TAO
2. Contextualiser pour le traitement automatique
3. Contextualiser pour le traducteur
4. Quels bons contextes pour le traducteur
   - Le bon exemple en lexicographie
   - Le contexte riche en connaissances en terminologie
   - Evaluer les besoins des traducteurs

Fournir des informations contextuelles au traducteur

- Buts :
  - Étoffer la liste des traductions possibles par des informations contextuelles
    - Catégorie grammaticale, cooccurrents de chaque *t'*
  - Associer à chaque traduction le passage illustrant au mieux son emploi :
    - Régularités, patrons linguistiques permettant de distinguer des sens ou des usages
    - Même métrique : tri des passages contenant *t'* en fonction d'une métrique exploitant les scores d'association des mots apparaissant dans le passage

Présentation des résultats

adjuvant (en) — fr

1. substitutif
2. adjuvant
3. conservateur
4. séquelle
5. hormonal
6. arsenal
7. néoadjuver
8. locorégiona
9. initial
10. systémique
11. bénéfier

Avec des informations contextuelles

Fr	1. substitutif	(A)	*hormonal, traitement*
	2. adjuvant	(N)	*chimiothérapie, traitement*
	3. conservateur	(N)	*traitement, séquelle*
	4. séquelle	(N)	*conservateur, traitement*
	5. hormonal	(A)	*récepteur, traitement*
	6. arsenal	(N)	*thérapeutique, partiel*
	7. néoadjuver	(V)	*traitement, superviser*
	8. locorégional	(A)	*récidive, traitement*
	9. initial	(A)	*traitement, bilan, tumeur*
	10. systémique	(A)	*traitement, adjuvant*
	11. bénéfier	(V)	*pouvoir, patiente, traitement*

Sélection d’un passage

adjuvant (en) fr 2. adjuvant

- Premier passage par ordre d’occurrence :

  Le risque de traitement de patientes en fait non éligibles du fait du statut HER2 non amplifié de leur tumeur est à prendre particulièrement en considération lorsqu’il doit s'appliquer en situation **adjuvante**, donc pour des patientes dont le pronostic est relativement bon, cela du fait que l'Herceptin est dotée d'une toxicité cardiaque et aussi pour des raisons économiques du fait du coût du traitement.

- Meilleur passage :

  Une patiente a eu une chimiothérapie première, 16 une chimiothérapie **adjuvante** et une chimiothérapie première et adjuvante.

adjuvant (en) fr 2. substitutif

- Premier passage :

Dans notre série, les patientes étaient pour la majorité préménopausées : dans 47,5 % des cas, dans 36,3 % ménopausées et dans 16,3 % ménopausées avec un traitement hormonal **substitutif**.

- Meilleur passage :

2. Bases et définition du traitement hormonal **substitutif** (THS)

Plan

1. Présentation du projet CRISTAL
   - Organisation et objectifs
   - La TAO
   - La notion de contexte
   - Utilisation de corpus comparables en TAO
2. Contextualiser pour le traitement automatique
3. Contextualiser pour le traducteur
4. Quels bons contextes pour le traducteur
   - Le bon exemple en lexicographie
   - Le contexte riche en connaissances en terminologie
   - Evaluer les besoins des traducteurs

Le bon exemple en lexicographie

- **Typicité** : des contextes typiques d'emploi des mots, des constructions fréquentes, des voisinages habituels

"clarify meaning, illustrate a word's contextual and combinatorial behaviour, and serve as models for language production " (Rundell & Kilgarriff, 2011)

"Providing examples of usage is an excellent means of helping the translator not only to select an appropriate equivalent but also to integrate it properly into a sentence. […] What translators are looking for is a variety of examples to illustrate different uses of the headword and its translation equivalents" (Roberts 1994)

Les principales fonctions d'un exemple lexicographique

- **Fonction syntagmatique ou distributionnelle** : composante idiomatique du lexique : constructions et collocations les plus usuelles

  Le mistral faisait voltiger nos manteaux et nous glaçait

- **Fonction paradigmatique** : synonymes, antonymes, hyperonymes, hyponymes, co-hyponymes

  Par quoi voudriez-vous que Virgile terminât ses hexamètres sinon par un dactyle et un spondée?

- **Fonction métalinguistique ou épilinguistique**

  L'hyperbate diffère de l'inversion (...) Cette figure (...) comprend l'anastrophe, la parenthèse et la synchyse

Bons et mauvais exemples

- Rey-Debove 2005

  « A vrai dire, toute phrase qui contient le mot-entrée peut servir d'exemple ; pour *chou* : *En faisant ses courses, il avait oublié le chou.* C'est néanmoins un mauvais exemple que personne n'aurait l'idée de forger. Car si on ne connaît pas le sens du mot *chou*, tout un paradigme de noms masculins est possible, y compris le *vinaigre*, le *pain*, le *savon*, le *journal*, etc. La seule information donnée est « chose qui s'achète » (*courses*) : ce qui exclut, par exemple, l'*ambition*, le *contrat*, etc. On appelle cela un mauvais exemple. A l'opposé, certains exemples sont comme des phrases à trous où le mot illustré manifeste l'essentiel de sa signification ; à *diluer* : *Cette peinture est trop épaisse, il faut la diluer avec de l'huile de lin* (DFC) ; à *prédire* : *Les prophètes prédirent la venue du Messie* ; à *fourreur : Acheter un renard, un manteau de vison chez un fourreur* (PR). On appelle cela un bon exemple, parce qu'on peut deviner la signification du mot d'après le contexte. »

Extraire automatiquement des bons exemples

- But : assister la sélection manuelle d'exemples dans un corpus :
- "Finding good examples in a mass of corpus data is labour-intensive. For all sorts of reasons, a majority of corpus sentences will not be suitable as they stand, so the lexicographer must either search out the best ones or modify corpus sentences which are promising but in some way flawed" (Rundell & Kilgarriff, 2011)
- GDEX (Good Dictionary Examples)
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P., 2008. GDEX: Automatically finding good dictionary examples in a corpus. In proceedings of XIII EURALEX Congress, E. Bernal & J. DeCesaris (Eds), pp.425-431, Barcelona, Universitat Pompeu Fabra.

Critères

- **Lisibilité :**
  - Bonne longueur
  - Mots suffisamment fréquents
  - Eviter les formes anaphoriques
  - Eviter la présence de caractères non alphabétiques
  - Préférer la présence du terme dans une proposition principale
- **Informativité :**
  - Favorise les phrases qui contiennent des mots qui sont souvent dans le voisinage du mot à traduire

GDEX

score (n;v)

object	goal :	A young Michael Owen scoring a goal then showing he 's
	hat-trick :	Round five England 43 - 22 Scotland Jamie Noon scored a
	equaliser :	A long pass through to Cook gave him the space he neec
	victory :	Ahead of all this , Lindley scored a fine victory , extendi
	try :	He was playing against the Baa-Baas and he scored a gre
	point :	The object is to become a wizard , a level attained on s
object_of	assign :	Each positive test assigns a weighted score to the incom
	achieve :	He is motivated to achieve high scores - even aiming for
a_modifier	overall :	The overall zetoc usage score in Table 2 is 7.8 .
	orchestral :	In 1953 J. Arthur Rank commissioned Whettam to write t
	musical :	To make the publication more helpful each song is prece
	final :	He also needs to be told his final score .
	average :	The average score for the librarians in the sample was 5.
	maximum :	Do the tests on the arm and leg on both sides of your b
modifies	sheet :	Ryman Premier top scorer Lee Boylan was on the score
n_modifier	credit :	You know what 's on your credit report , but would you
	autograph :	Sketches , fragments and fair copies of compositions and
	deprivation :	In addition to a wide range of general statistics , the dep
pp_with	volley :	The intense pressure saw Musson cross for Gibson to sc
	header :	Hreidarsson opens the scoring with a header from a corı

Les contextes riches en connaissance

- Notion liée au domaine de la gestion des connaissances :
  - Construction de bases de données terminologiques et d'ontologies

Ingrid Meyer 2001 :

« By knowledge-rich context, we designate a context indicating at least one item of domain knowledge that could be useful for conceptual analysis »

- Présence de marqueurs de relations conceptuelles
- *[**méronymie**] Un **volcan** se compose de trois parties : un **réservoir**, une **cheminée** et un **édifice** visible en surface.*
- *[**hyperonymie**] There are three types of **lava** and **lava flows** : **pillow**, **pahoehoe**, and **aa**.*

Extraction de patrons marquant les relations conceptuelles

- Acquisition semi-automatisée à partir de textes de patrons lexico-syntaxiques marquant une relation donnée

  Hearst 1992 et 1998 : méronymie, hyperonymie

- Exemple :

NP_0 *such as* NP_1 {,NP_2 ... , (*and* | *or*) NP_i} i ≥1

=>

pour tout NP_i, i ≥ 1, hyponyme(NP_i, NP_0)

- (S1) Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

hyponyme(“Gelidium”,“red algae”).

- Méthode reprise par (Morin 1999), (Condamines et Rebeyrolle 2001)
- Plus récemment : (Pantel & Pennachiotti 2008)

Principe (Morin 1999 pour l'hyperonymie)

1) Choix d'une amorce : thésaurus d'agronomie
   - Ex : *glycérol* EST-UN *polyol*
2) Extraction des phrases contenant le couple
   - Ex : *l'hydrolyse est activée par le glucose et les **polyols** tels que le sorbitol et le **glycérol***
3) Identification d'un environnement commun qui généralise les phrases extraites = **schéma lexico-syntaxique candidat**
   - Ex : NP tel que LISTE
4) Validation des schémas candidats les plus pertinents
6) Extraction de nouveaux couples candidats à l'aide de ces nouveaux schémas
7) Validation des couples. Réitération du processus à partir de 3)

Exemples

- SN, particulièrement SN

« … des espèces de phlébotomes anthropophiles, particulièrement Lutzomyia trapidoi »

- SN {et|ou} de autre SN

« issues des deux blés ou d'autres céréales »

- SN tel LISTE

« Les caractéristiques du site telles que la pente, le sous-bois et la distance des usines … »

- SN et notamment SN

« … le développement de bactéries et notamment de germes lactiques »

- Chez le SN, SN

« Chez les Phalaenopsis, Orchidées monopodiales, … »

Dans le cadre du projet

- Constitution d’un répertoire de CRC à partir de travaux précédents
  - En français et en anglais
- Complété à partir des corpus de travail du projet
- Prise en compte de la variation en genre et domaine

Quels sont les bons contextes pour le traducteur ?

- Nombreuses études réalisées pour comprendre le fonctionnement des traducteurs et améliorer les outils et l'enseignement de la traductologie :
  - Quelles ressources d'information
    - Dictionnaires monolingues et bilingues, généraux et spécialisés
    - Corpus de textes cibles, corpus alignés
  - Comment sont-elles utilisées
- Constats :
  - Les traducteurs cherchent à résoudre des problèmes très dépendants du contexte (Varantola 1998)
  - Double insuffisance des dictionnaires :
    - Peu de contextes
    - Des unités lexicales courtes

    What they typically find in term banks are definitions and terms presented out of context, or in only one single context" (Bowker, 2011).

Une très large gamme de besoins

(Bowker 2012)

- Des informations de type spécialisé et de type général
- Des données chiffrées relatives à la fréquence (implantation)
- Des informations sur les relations lexicales et conceptuelles
  - Relations syntagmatiques et paradigmatiq
- Des informations sur l'usage
  - Collocations (collocates clouds)
- Des informations sur le style
  - Genre, registre
- Autres : aspects pragmatiques, *tricky translations*, usages à éviter …

Evaluer l'intérêt des contextes pour le traducteur : expérience à venir

- 3 sites : Toulouse (CETIM), Genève, Le Mans
  - Prétest en décembre à Genève
  - Expérience en janvier 2014 à Toulouse
- Principes :
  - Un texte à traduire (volcanologie) comportant des termes
  - Un lexique *ad hoc*
  - Une sélection de contextes en langue source :
    - Mêlant contextes riches et contextes pauvres
  - Une sélection de contextes en langue cible ?
  - Observation des comportements

Conclusion

- Etoffer la notion de bons contextes :
  - À partir des travaux réalisés dans des disciplines connexes
    - Notion élargie de contexte riche en connaissance pour le traducteur
  - En évaluant leur apport pour les traducteurs
  - En menant des analyses à partir des corpus de travail et des contextes produits automatiquement
- Intégrer ces contextes enrichis :
  - Dans la procédure automatique d'identification des traductions
  - Dans la plate-forme de TAO

Références

- Altenberg, B. & Granger, S. (2002). Recent Trends in Cross-Linguistic Lexical Studies. In B. Altenberg & S. Granger (Eds.), *Lexis in Contrast*, *Corpus-Based Approaches* (pp. 3-48). Amsterdam, Philadelphie: John Benjamins.
- Blancafort, H., Heid, U., Gornostay, T., Méchoulam, C., Daille, B. & Sharoff, S., (2011). User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT Tools. In *In Tralogy, Session 1 - Terminologie et Traduction. Paris, France.*
- Bowker, L. & Pearson, J., (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London/New York: Routeledge.
- Bowker L., 2011. “Off the record and on the fly,” *Corpus-based Translation Studies: Research and Applications* (Eds. A. Kruger, K. Wallmach and J. Munday). London/New York: Continuum, pp. 211-236.
- Bowker L., 2012. “Meeting the needs of translators in the age of e-lexicography: Exploring the possibilities”, in . S. Granger & M. Paquot (eds) *Electronic Lexicography*, Oxford University Press, pp. 379-387.
- Condamines, A. & Rebeyrolle, J. (2001). Searching for and Identifying Conceptual Relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB) : method and results (pp.127-48). In D. Bourigault, M.-C. L’homme & C. Jacquemin, (Eds.), *Recent Advances in Computational Terminology*. Amsterdam, Philadelphia: John Benjamins.

- Delpech, E., Daille, B., Morin, E. et Lemaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora: compositional translation and ranking. Proceedings, *24th International Conference on Computational Linguistics (COLING 2012)*, pp. 745-761, Mumbai, India. 2012.
- Harastani, R.; Daille, B. & Morin, E. (2013), Ranking Translation Candidates Acquired from Comparable Corpora, *in 'Proceedings of the Sixth International Joint Conference on Natural Language Processing', Asian Federation of Natural Language Processing, Nagoya, Japan, pp. 401--409.*
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P., 2008. GDEX: Automatically finding good dictionary examples in a corpus. In proceedings of XIII EURALEX Congress, E. Bernal & J. DeCesaris (Eds), pp.425-431, Barcelona, Universitat Pompeu Fabra.
- Li, B., et Gaussier, E.. "Improving corpus comparability for bilingual lexicon extraction from comparable corpora." *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010.

- Morin, E. et Daille, B. (2012). Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. *Actes de la conférence conjointe JEP-TALN-RECITAL*, Grenoble.
- Roberts, Roda P., 1994. *Bilingual Dictionaries Prepared in Terms of Translators' Needs*, Proceedings of CTIC 3rd Conference (May 4-8, 1994), Translation in the Global Village, Banff, CTIC, pp. 51-65.
- Varantola, K., 1998. Translators and their Use of Dictionaries: User Needs and User Habits.