Dictionnaires/glossaires vs corpus : ce que les corpus font (ou ne font pas) à la terminologie pour traducteurs

Laurent Gautier, professeur des Universités, Centre Interlangues Texte Image Langage (UBFC, EA 4182)

Structure de la présentation

1. Problématique et objectifs
2. Le renouveau de la terminologie
3. Corpus et (apprentissage de la) traduction
4. Vers une terminologie de corpus
5. Conclusion et perspectives

1. Problématique et objectifs

Le paradigme du traducteur technologue

- Entendu à l'atelier EMT2013 (13.09.2013) :
  - Inguna Skadiņa: modern translator has to be a **terminologist**, editor, project manager, **tech support expert**....
  - Inguna Skadiņa: 3 courses needed for translators: **language technologies; CAT tools; human-assited translation**

⇒**Point commun : la dimension technologique de la traduction s'accroit et ne limite pas à l'environnement du traducteur, elle modifie aussi notre conception de la langue et de la terminologie**

- Compétences en extraction d'information :
  terminologie
  recherche documentaire,
  logiciels de terminologie
- Compétences technologiques :
  outils d'aide à la traduction
  gestion de bases de données
  création et gestion de corpus
- Compétences thématiques :
  connaissances dans des domaines de spécialité et matières d'application

=> Dénominateurs communs : la langue en usage, interrogeable/ requêtable, est documentée dans les corpus

Vers une compétence croisée et holistique ?

- Terminologie : rôle historique à la fois dans la formation et dans la pratique des traducteurs :
  - Extraction manuelle, puis semi-automatisée
  - Rôle des experts-valideurs
  - Rôle « fétiche » du « glossaire de traduction » / »dictionnaire spécialisé » puis de la BD terminologique
- Quelles évolutions possibles à l'heure des nouveaux environnements du traducteur ?
  - Terminologie *in vivo* (vs. *in vitro*)
  - Porte d'entrée vers les contenus spécialisés (compréhension sémantique) *et* la mise en discours (choix de formulation)
  - Apports des masses de données textuelles existantes

2. Le renouveau de la terminologie

Un tournant cognitif *aussi* en terminologie … ?

- Dépassement de la perspective objectiviste de la terminologie/nomenclature wüsterienne reconnu depuis la socio-terminologie de Gaudin (1993)
- Évolution qui, comme celle des LSP, est indexée sur l'évolution de la recherche en sciences du langage :
    - Tournant pragmatique : terminologie située, utilisée et vécue
    - Tournant discursif : mode d'emploi textuel du terme
    - Tournant cognitif : le terme comme segment de connaissance indissociable d'une architecture des savoirs

- Les catégories traditionnelles du terme (aujourd'hui largement remise en cause):
  - Univocité
  - Non-ambigüité
  - Objectivité / neutralité
  - Non-synonymie
- aujourd'hui place reconnue à :
  - la variation
  - la dimension cognitive
  - au rôle fondamental des co- et contextes

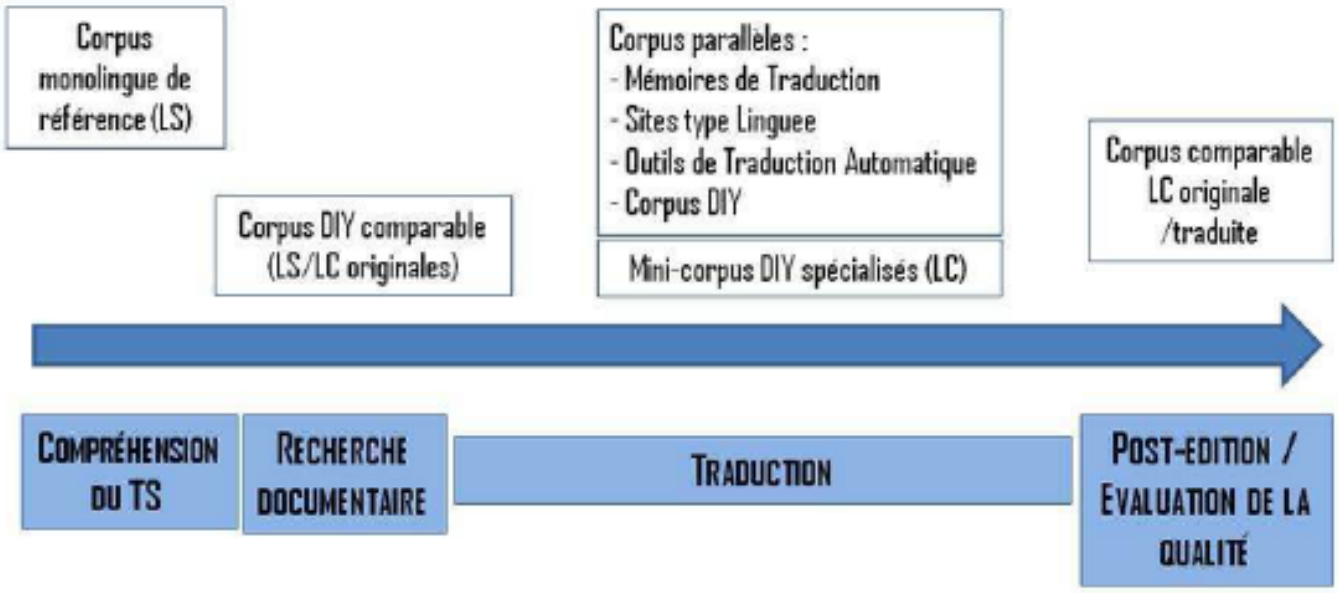
  => Uniquement observable *in vivo*, donc en corpus!

3. Corpus et (apprentissage de la) traduction

Prémisses

- Omniprésence explicite ou non des corpus dans les outils de TAO et environnement du traducteur : MT, TA, BD termino, sites grand public (*linguee*)
- Définition de base : ensemble de données langagières (écrites et / ou orales) organisées, répondant à un objectif d'exploitation et remplissant un certain nombre de conditions. Ces données sont ensuite préparées pour donner lieu à des traitements automatiques plus ou moins poussés
- Critère d'**authenticité** : donné la plupart du temps, mais à traiter avec précaution quand le corpus doit être recueilli
- Critère de **représentativité**, garant de la validité des résultats

- Conditions essentielles :
  - En adéquation totale avec le **besoin de traduction**, en lien avec le flux du projet (Loock 2017 : 19):



Typologie générale

- Corpus monolingues « dits » de référence, en *monitoring*
- Un exemple pour l'anglais : https://corpus.byu.edu/coca/

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, and the only large and balanced corpus of American English. COCA is probably the most widely-used corpus of English, and it is related to many other corpora of English that we have created, which offer unparalleled insight into variation in English.The corpus contains more than **560 million words of text (20 million words each year 1990-2017)** and it is equally divided among spoken, fiction, **popular magazines, newspapers, and academic texts.**

- Corpus plurilingues
  - Corpus issus de projets de recherche : http://genealogiesofknowledge.net/translational-english-corpus-tec/
  - Corpus issus d'institutions : https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

The Acquis Communautaire (AC) is the total body of European Union (EU) law applicable in the the EU Member States. This collection of legislative text changes continuously and currently comprises selected texts written between the 1950s and now. As of the beginning of the year 2007, the EU had 27 Member States and 23 official languages. The Acquis Communautaire texts exist in these languages, although Irish translations are not currently available. The Acquis Communautaire thus is **a collection of parallel texts** in the following 22 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovenian and Swedish.

  - Corpus *ad hoc* (DIY-Corpus)

  => Dimension capitale pour le traducteur : parallèles ou comparables / langue originale ou traduite

Corpus parallèles

- « A ‘parallel corpus’ is a bilingual or multilingual corpus that contains one set of texts in two or more languages. » (Teubert 1996 : 245)
  - textes traduits
  - *tertium comparationis2* : équivalence supposée entre les textes
  - préparation nécessitant une opération d’alignement (semi-)automatique

=> LF Aligner https://sourceforge.net/p/aligner/wiki/Home/

- Utilisations principales :
  - recherche d’équivalents terminologiques : domaine plus ou moins autonome en terminologie et pour les principaux outils de TAO (ATR)
  - recherche de collocations / combinatoires
  - recherche de traits stylistiques / patterns

Alignement de segments sur un extrait contenant le terme "growth"

Anglais	Français	Allemand	Néerlandais
Given high structural unemployment and low potential output growth in the euro area, a cyclical recovery along the lines of the March ECB staff projections is no grounds for complacency.	Au vu du niveau élevé de chômage structurel et de la faible croissance potentielle dans la zone euro, une reprise conjoncturelle telle que celle ressortant des projections de mars des services de la BCE **ne permet aucun excès de confiance.**	Angesichts der hohen strukturellen Arbeitslosigkeit und des geringen Wachstums des Produktionspotenzials im Eurogebiet **gibt** eine Konjunkturerholung wie in den von Experten der EZB erstellten Projektionen vom März **keinen Anlass zur Sorglosigkeit.**	Gezien de hoge structurele werkloosheid en de lage potentiële productiegroei in het eurogebied, **is** een conjunctuurgebonden herstel zoals geschetst in de door medewerkers van de ECB opgestelde projecties van maart **geen reden om achterover te leunen.**

Corpus comparables

- « ’Comparable corpora’ are corpora in two or more languages with the same or similar composition. All copora have an explicit or implicit composition. The texts they contain can be classified according to a variety of intralinguistic or extralinguistic features. » (Teubert 1996 : 245)
- *Tertium comparationis* : à définir en fonction des objectifs, seul garant de l’homogénéité du corpus

→ un tc inscrit au niveau textuel / discursif

=> Document BoE / BnF

- TC possibles :
  - un type de texte
  - un domaine thématique
  - une situation énonciative
- Apports essentiels :
  - analyse de deux langues « originales »
  - travail aux niveaux textuel et discursif
  - travail possible au niveau microlinguistique (ATR par exemple) mais sans appariement automatique

4. Vers une terminologie de corpus

Approche intuitive

- Corpus *adhoc* français construit avec WebBootCat sous Sketchengine
- Ce qu'apprend une grammaire lexicale locale

avalanche *(noun)*
avalanches freq = 3,630 (5,233.36 per million)

verbs with "avalanche" as object		11.68
déclencher +	148	12.69
déclencher une avalanche		
provoquer	22	10.41
provoquer une avalanche		
une	12	9.64
Une avalanche		
produire	10	9.39
enfouir	7	8.92
classer	6	8.82
survenir	6	8.79
prévenir	6	8.70
éviter	6	8.57
recenser	5	8.53
risquer	5	8.53
souiller	4	8.25
dévier	4	8.25
simuler	4	8.24
ces	4	8.19
voir	6	8.13
choisir	4	8.05
représenter	4	8.02
avoir	15	7.91

verbs with "avalanche" as subject		3.55
pouvoir	46	11.09
avalanche peut		
partir	7	10.54
emporter	5	10.22
concerner	5	10.07
venir	5	9.93
constituer	4	9.82
tuer	3	9.53
couper	3	9.51
impliquer	3	9.48
évoluer	3	9.34
aller	4	9.07
faire	4	9.00
rester	3	8.98

modifiers of "avalanche"		8.43
spontané	22	10.91
avalanches spontanées		
gros	27	10.73
grosses avalanches		
accidentel	15	10.53
% des avalanches accidentelles		
naturel	17	10.22
les avalanches naturelles		
coulant	11	10.13
avalanche coulante		
dense	9	9.77
petit	16	9.56
petites avalanches peuvent se produire		
poudreux	8	9.48
mortel	7	9.39
nouveau	12	9.37
nouvelles avalanches		
préventif	8	9.31
spectaculaire	6	9.27
mixte	6	9.23
nombreux	9	9.19
possible	7	9.17

adjective predicates of "avalanche"		1.27
endémique	3	10.97
fort	4	10.79
fréquent	3	10.77
important	4	10.44

"avalanche" is a ...		0.88
masse	7	12.51
utilisation	3	11.45
technique	3	11.34

prepositional phrases		
"avalanche" de	462	12.73
"avalanche" en	63	1.74
"avalanche" à	49	1.35
"avalanche" du	22	0.61
"avalanche" au	19	0.52
"avalanche" pour	15	0.41
"avalanche" par	10	0.28
"avalanche" avec	9	0.25
"avalanche" sans	3	0.08

"avalanche" and/or ...		5.59
taille	5	9.53
dva	4	9.16
sonde	5	9.15
risque	4	8.90
deglace	3	8.90
alpug	3	8.90
schweizerische	3	8.90
cemagref	3	8.89
pida	3	8.88
donnée	3	8.88
imprimerie	3	8.87
avalanche	6	8.85
prix	3	8.85
prévision	3	8.82
secours	4	8.79
neige	5	8.79
chute	3	8.71
sécurité	3	8.45

less "avalanche" than ...		0.11
volumineux	4	13.99

Du dictionnaire au corpus :

- « (Ce) sont de véritables dictionnaires de nouvelle génération (qui) exploitent des corpus regroupant des traductions segmentées généralement au niveau de la phrase. » (Loock 2016 : 35)

Limites :

- « Si la fiabilité des traductions pose parfois problème, il s'agit bien de proposer un échantillon extrait de corpus parallèles (même si le statut langue source / langue cible n'est pas toujours évident) collectés de façon automatique par le biais d'un robot, corpus au sein desquels est sélectionnée selon un algorithme une série de paires d'exemples pour les deux langues demandées. » (Loock 2016 : 35)

- Le corpus transforme l'approche de la terminologie pour intégrer le terme dans des segments de connaissances spécialisées « prêtes à traduire » :

  « The typical linguistic features of ESP **cannot be characterised as a list of discreet items** (technical terminology, the passive, hedging, impersonal expressions, etc.), rather the most typical features of ESP texts are **chains of meaningful interlocking lexical and grammatical structures**, wich we have called **lexico-grammatical patterns**. » (Gledhill/Kübler 2016 : 75)

  =>Les grammaires de construction et la suspension de la dichotomie lexique vs. grammaire

Alignement de segments sur un extrait contenant le terme "growth"

Anglais	Français	Allemand	Néerlandais
Given high structural unemployment and low potential output growth in the euro area, a cyclical recovery along the lines of the March ECB staff projections is no grounds for complacency.	Au vu du niveau élevé de chômage structurel et de la faible croissance potentielle dans la zone euro, une reprise conjoncturelle telle que celle ressortant des projections de mars des services de la BCE **ne permet aucun excès de confiance.**	Angesichts der hohen strukturellen Arbeitslosigkeit und des geringen Wachstums des Produktionspotenzials im Eurogebiet **gibt** eine Konjunkturerholung wie in den von Experten der EZB erstellten Projektionen vom März **keinen Anlass zur Sorglosigkeit.**	Gezien de hoge structurele werkloosheid en de lage potentiële productiegroei in het eurogebied, **is** een conjunctuurgebonden herstel zoals geschetst in de door medewerkers van de ECB opgestelde projecties van maart **geen reden om achterover te leunen.**

- « Termes » traditionnels = nœuds de structures conceptuelles organisant le domaine cognitif de référence « traduits » en langue par des mises en mots récurrentes analysables en chaînes prédicat-arguments

  La syntaxe formelle limite la distribution des mots sur la base de schémas de bonne formation des expressions complexes qui ne sont pas sensibles au contenu. La syntaxe des concepts limite la distribution des mots sur **la base de procès et d'états de choses cohérents et cognitivement adéquats**, qui fournissent à la description du contenu des mots une dimension relationnelle supplémentaire. (Prandi 1998 : 36, souligné par nous)

- Figement réinvesti en termes de **récurrences** – **fossilisation => stabilité** (Feilke 1996) – **formulations préférentielles,** puis **exclusives**
- Saisie du texte spécialisé et du domaine de spécialité comme concaténations de **répertoires restreints** : lexique/terminologie + contraintes de linéarisation + codes grammaticaux

5. Bilan et perspectives

- L'approche par corpus a révolutionné, pour les discours spécialisés, la conceptualisation même de la terminologie, garante d'une **idiomaticité de corpus**

- Vers un **lexique-grammaire** :
  - Intégrant la terminologie dans des *structures prédicat-arguments* reflets de la grammaire conceptuelle du champ ;
  - Listant les répertoires de *formes morpho-syntaxiques* préférentielles associées ;
  - Associant le lexique de liaison vu aussi en termes de fossilisation.

- En cohérence avec l'approche holistique des discours spécialisés
- Bouleversement des outils mêmes de représentation des discours spécialisés (en particulier bases de données qui deviennent des **bases d'organisation de l'information**)

Merci pour votre attention !

Laurent Gautier (laurent.gautier@ubfc.fr)