April 12, 2017

Simplification et normalisation en traduction: Evaluation d'une prédiction à propos de l'emploi des collocations par l'analyse automatique d'un corpus parallèle et comparable

Yves Bestgen

1er Congrès mondial de Traductologie

Université Paris-Nanterre

Mercredi 12 avril 2017

Simplification et normalisation en traduction

Evaluation d'une prédiction à propos de l'emploi des collocations par l'analyse automatique d'un corpus parallèle et comparable

Yves Bestgen

Centre for English Corpus Linguistics
Université catholique de Louvain

Plan

- Cadre théorique
  - Traduction et phraséologie
  - Méthodologie pour l'analyse des collocations
- Analyse de corpus
  - Hypothèse
  - Corpus
  - Méthode
  - Analyses et résultats
- Discussion et conclusion

Traduction et phraséologie

- A la recherche d'éventuels universaux de traduction

  Invariant features which characterize all translated texts independently of source language and translation direction

  (Zanettin 2013, p.21; M. Baker 1993, 2007)

  - Simplification
  - Normalisation
  - Explicitation
  - ...

- Thèse controversée

  (Becher 2010; Loock 2012; Mauranen & Kujamäki 2004)

Traduction et **phraséologie**

- Importance des manières conventionnelles de s'exprimer

  (Cowie 1994; Sinclair 1991)

- Différents types d'unités dont
  - Collocations : cooccurrences privilégiées
    - Deux mots qui s'observent l'un près de l'autre plus souvent que le hasard ne le prédit (Sinclair 1991)
    - Les analyses porteront sur les mots contigus (bigrammes)

      inversement proportionnel, avoir besoin, par exemple

- Importante conséquence en situation multilingue (Colson 2008)
  - Spécifiques à une langue

Traduction et phraséologie

- Normalisation et simplification dans les textes traduits
  - Préférence pour les collocations usuelles dans la langue cible au détriment des collocations rares

    (M. Baker 2004, 2007; Laviosa 2004)
  - Arguments empiriques issus de l'analyse manuelle approfondie d'un nombre limité de collocations
    - Résultats contradictoires

      (Dayrell 2007; Kenny 2001; Marco 2009...)

Méthodologie pour l'analyse des collocations

- Normalisation, simplification et phraséologie
  - Préférence pour les collocations usuelles dans la langue cible au détriment des collocations rares

- Tester cette prédiction par une analyse automatique?
  - Identifier automatiquement les collocations usuelles et les collocations rares dans un texte traduit ou non-traduit?

Méthodologie pour l'analyse des collocations

- Identifier automatiquement les collocations usuelles et les collocations rares?

- Sur la base d'un corpus de référence de la langue cible
- La simple fréquence de cooccurrence ne suffit pas
- Nombreux indices d'association collocationnelle (Evert 2008)
- Deux indices sont particulièrement pertinents
  - *t* : privilégie les collocations plus fréquentes
    - *avoir besoin*, *par exemple*
  - *Information mutuelle (IM)*: privilégie les collocations plus rares
    - *violation flagrante, inversement proportionnel*

(P. Baker 2006; Church *et al.* 1991; Hunston 2002)

Méthodologie pour l'analyse des collocations

- Travaux antérieurs
  - En traductologie
    - Bernardini 2007
    - Ferraresi *et al.* 2015
    - (Volansky et al. 2013)
  - En apprentissage des langues étrangères
    - Durrant & Schmitt 2009
    - Bestgen & Granger 2014

Plan

- Cadre théorique
  - Traduction et phraséologie
  - Méthodologie pour l'analyse des collocations
- Analyse de corpus
  - Hypothèse
  - Corpus
  - Méthode
  - Analyses et résultats
- Discussion et conclusion

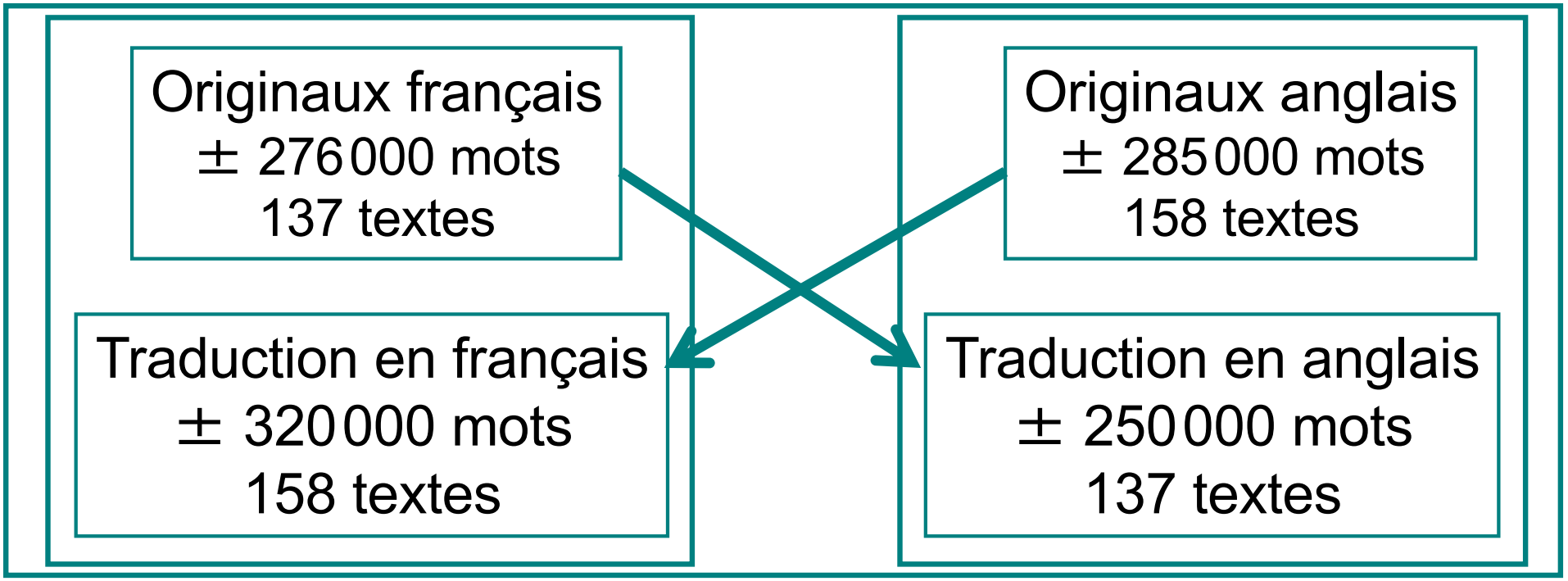
Hypothèse

- Normalisation, simplification et phraséologie
  - Préférence pour les collocations usuelles dans la langue cible au détriment des collocations rares

- Analyser dans les textes le rapport entre
  - La proportion de collocations pour t
  - La proportion de collocations pour IM
- Ce rapport devrait être plus élevé dans un texte traduit

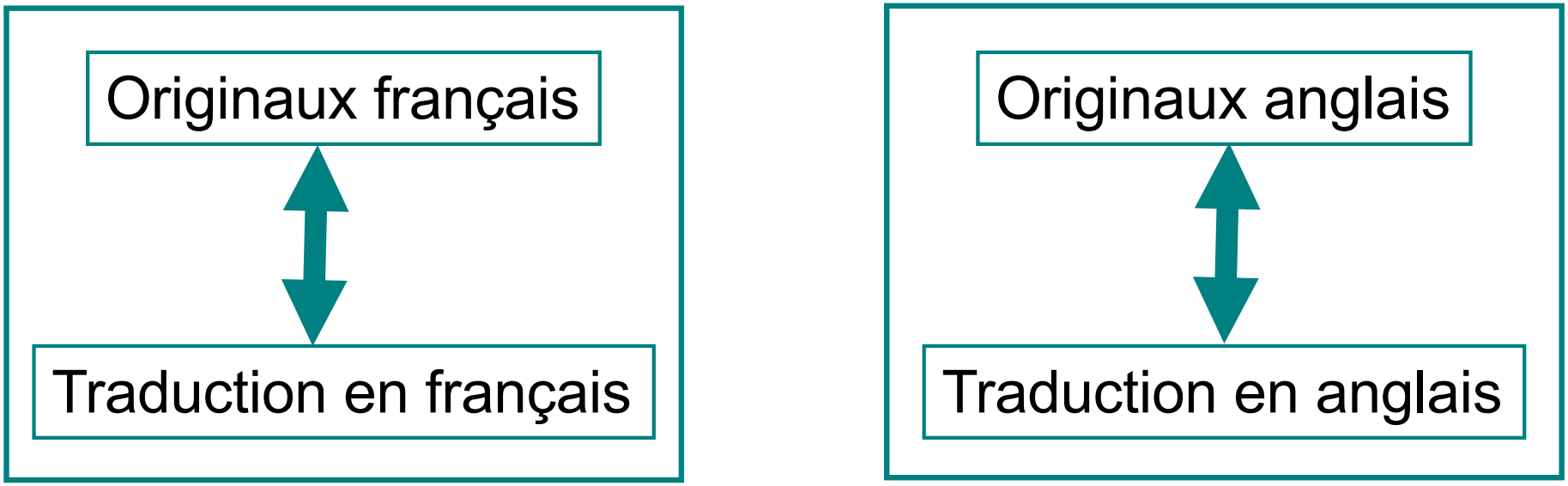
Corpus

- Corpus d'évaluation
  - PLECI (Poitiers-Louvain Échange de Corpus Informatisé)
    - Parallèle et comparable
    - Section : articles de journaux



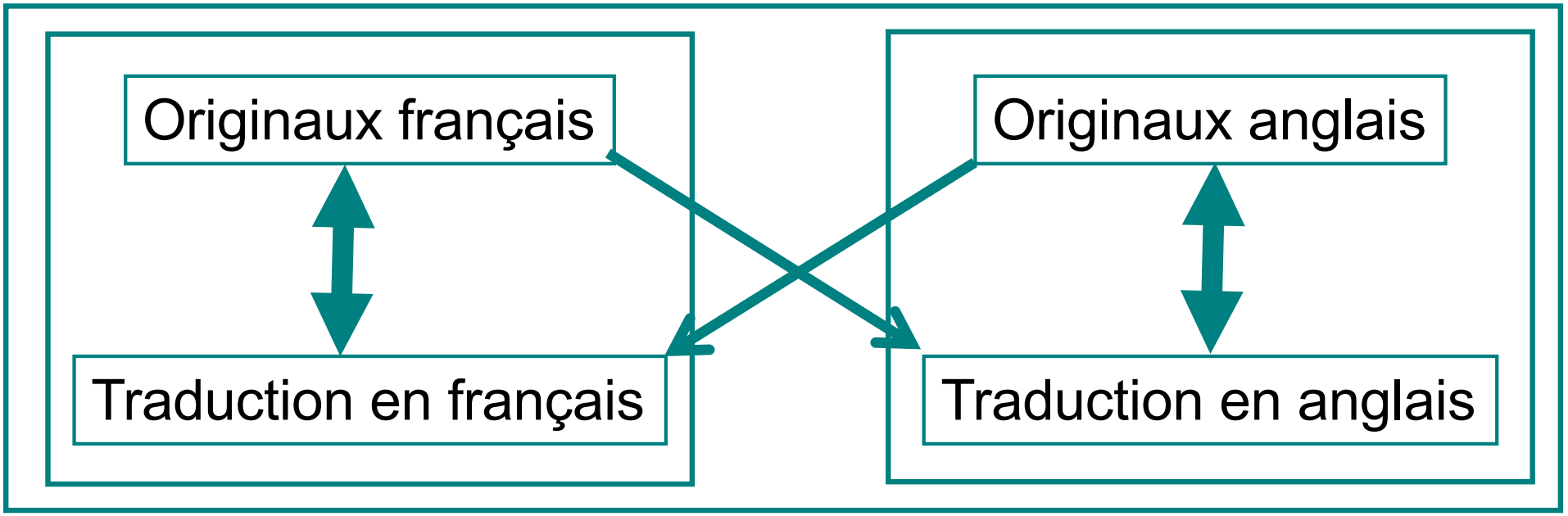
Corpus

- Double analyse de corpus parallèles et comparables



Corpus

- Double analyse de corpus parallèles et comparables
  - Permet un meilleur contrôle des différences entre les textes en langue source



Corpus

- Corpus d'évaluation
  - PLECI (Poitiers-Louvain Échange de Corpus Informatisé)
    - Parallèle et comparable
    - Section : articles de journaux
    - Prétraitement :
      - Segmenté en articles
      - Lemmatisation par TreeTagger (Schmid, 1994)

Corpus

- Corpus de référence
  - WaCKy (***W**eb **a**s **C**orpus **k**ool **y**nitiative* : Baroni *et al.* 2009)
    - *UkSubset* : 100 millions de mots
    - *FrSubset* : 100 millions de mots

Méthodologie pour l'analyse des collocations

- Score collocationnel d'un texte
  1. Extraction des bigrammes

Extraction des bigrammes

Mot	*LEMME*
Fuyant	
des	FUIR_DU
impôts	DU_IMPÔT
excessifs	IMPÔT_EXCESSIF
et	EXCESSIF_ET
une	ET_UN
administration	UN_ADMINISTRATION
tatillonne	ADMINISTRATION_TATILLON
,	
mais	
également	MAIS_ÉGALEMENT
poussés	ÉGALEMENT_POUSSER
par	POUSSER_PAR
le	PAR_LE
goût	LE_GOÛT
de	GOÛT_DE
l'	DE_LE
aventure	LE_AVENTURE

Méthodologie pour l'analyse des collocations

- Score collocationnel d'un texte
  1. Extraction des bigrammes
  2. Détermination de l'intensité collocationnelle

     Sur la base du corpus de référence

Détermination de l'intensité collocationnelle

Mot	*LEMME*	*IM*	*t*
Fuyant			
des	FUIR_DU	-0,8	-0,7
impôts	DU_IMPÔT	2,8	36,5
excessifs	IMPÔT_EXCESSIF	3,0	0,8
et	EXCESSIF_ET	1,4	5,5
une	ET_UN	0,0	9,2
administration	UN_ADMINISTRATION	0,1	1,7
tatillonne	ADMINISTRATION_TATILLON	9,1	2,2
,			
mais			
également	MAIS_ÉGALEMENT	4,8	50,8
poussés	ÉGALEMENT_POUSSER	-1,3	-1,4
par	POUSSER_PAR	2,8	13,3
le	PAR_LE	2,2	381,8
goût	LE_GOÛT	1,9	32,0
de	GOÛT_DE	1,5	21,7
l'	DE_LE	1,4	744,3
aventure	LE_AVENTURE	2,3	38,5

Méthodologie pour l'analyse des collocations

- Score collocationnel d'un texte
  1. Extraction des bigrammes
  2. Détermination de l'intensité collocationnelle
  3. Catégorisation selon l'intensité collocationnelle

	IM	t
Collocationnel	IM ≥ 5	t ≥ 6
Non-collocationnel	IM < 5	t < 6

(Durrant & Schmitt 2009)

Exemples de bigrammes collocationnels

- IM - français : *ordure ménager, réchauffement climatique, inversement proportionnel, tournant décisif, proprement dit*.
- IM - anglais : *mentally retarded, ballistic missile, double-edged sword, behave responsibly, grossly exaggerate*.
- t - français : *de le, ainsi que, afin de, en place, nombre de*.
- t - anglais : *of the, it be, part of, more than, such as*.

(bigrammes lemmatisés)

Méthodologie pour l'analyse des collocations

- Score collocationnel d'un texte
  1. Extraction des bigrammes
  2. Détermination de l'intensité collocationnelle
  3. Catégorisation selon l'intensité collocationnelle
  4. Obtention du score collocationnel du texte

	t		IM	
	NC	C	NC	C
Fréquence	418	1371	1546	243
Pourcentage	23,4	76,6	86,4	13,6

$$Rapport = \frac{76,6}{13,6} = 5,64$$

NC = non-collocationnel ; C = collocationnel ; (1371/243 = 5,64)

Analyses et résultats

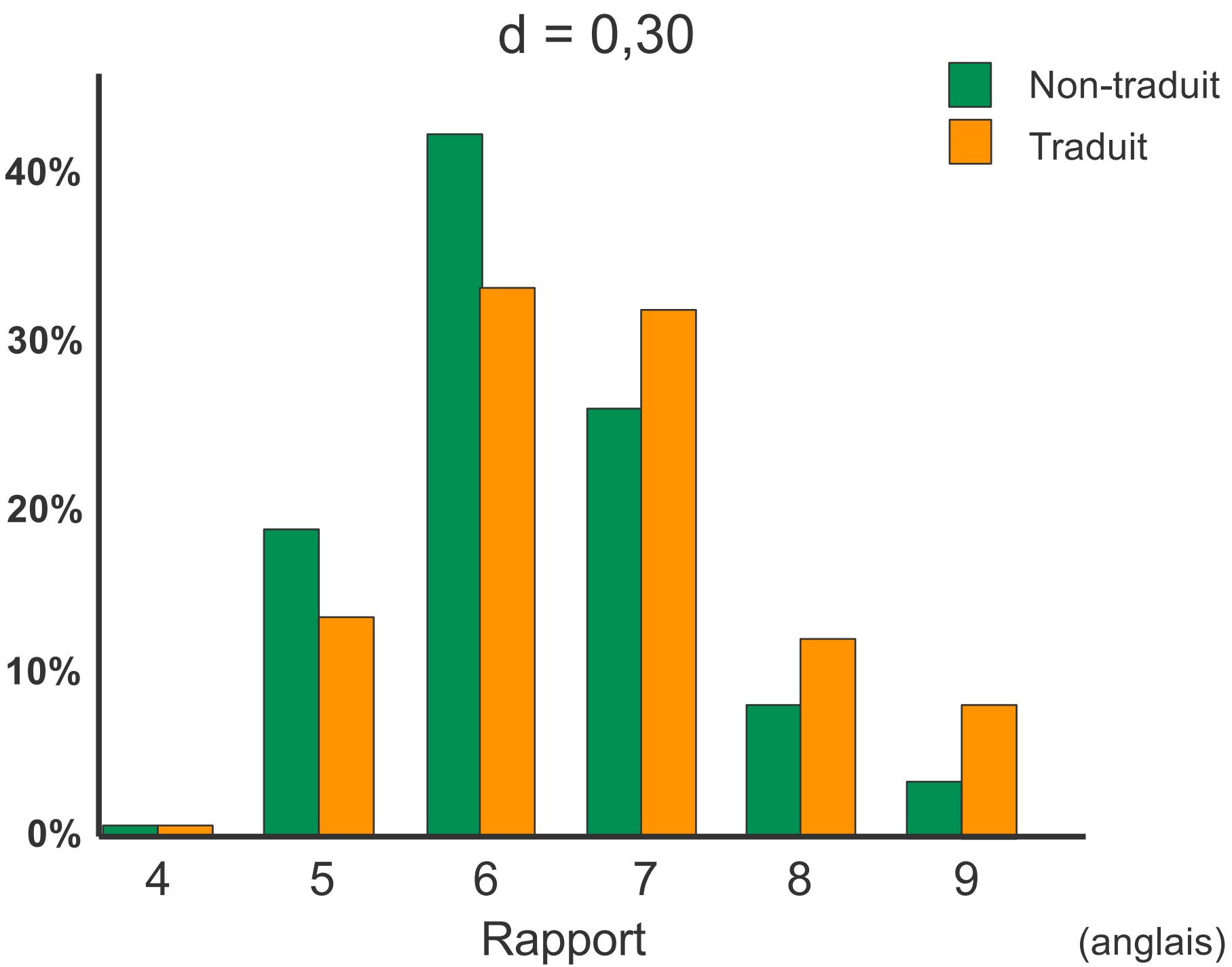
- Test statistique
  - Test de Student pour comparer des moyennes
    - Seuil de décision : $p \leq 0,05$
    - Taille de l'effet : *d* de Cohen
      - Différence entre les moyennes en fonction de la variabilité des scores
      - 0,20 = petit; 0,50 = moyen; 0,80 = grand

Résultats

- Le rapport collocationnel t / IM est-il plus élevé dans les textes traduits?

	Original	Traduit	Prob.	D de Cohen
Français	7,69	8,01	0,016	0,28
Anglais	6,36	6,72	0,008	0,31

- Oui, mais les tailles d'effets sont faibles


d = 0,30
Non-traduit
Traduit
40%
30%
20%
10%
0%
4
5
6
7
8
9
Rapport
(anglais)

Discussion et conclusion

- Synthèse
  - Une analyse globale automatisée des collocations

    (Bernardini, 2007; Durrant & Schmitt 2009; Bestgen & Granger 2014; Ferraresi et al. 2015)
  - Des résultats similaires dans les deux langues en accord avec l'hypothèse de normalisation–simplification
    - Les textes traduits présentent un rapport collocationnel t / IM plus élevé que les textes non traduits

Discussion et conclusion

- Limitations et développements
  - Un seul corpus d'évaluation :
    - Reproduire avec d'autres corpus d'évaluation, mais aussi d'autres genres de textes, avec plus de langues sources et pour d'autres langues cibles
  - Une analyse globale, automatisée et quantitative :
    - Analyser uniquement certains patterns syntaxiques
    - Analyse approfondie des bigrammes typiques des textes traduits vs. non-traduits
    - Retourner aux textes : comparer les deux versions

Merci de votre attention

Principales références

- Baker, M. (1993). Corpus linguistics and translation studies: implications and applications. In *Text and Technology*. Benjamins. 233-250.
- Baroni et al. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*, 209-226.
- Bernardini, S. (2007). Collocations in Translated Language. Combining parallel, comparable and reference corpora. Proceedings of the *Corpus Linguistics Conference*.
- Bestgen, Y & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 2014.
- Colson, J.-P. (2008). Cross-linguistic phraseological studies: An overview. In *Phraseology: An Interdisciplinary Perspective*. Benjamins. 191–206.
- Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL*, *47*, 157-177.
- Ferraresi, A., Bernardini, S. and Miličević, M., 2015. Collocations across languages: evidence from interpreting and translation. In: Corpus Linguistics 2015. Lancaster: UCREL, 106-108.
- Loock, R., 2012. La traductologie sur corpus : études de cas et enjeux. In: Actes du colloque Traduction et Traductologie. Mons: CIPA, 99-116.
- Zanettin, F. (2013). Corpus methods for descriptive translation studies. *Procedia - Social and Behavioral Sciences*, *95*, 20-32.