Université de Montréal
Département d'Informatique et de Recherche Opérationnelle
Laboratoire RALI

PROBLÉMATIQUE DE LA TRADUCTION DE BULLETINS ANGLAIS VERS L'INUKTITUT

Chiheb Trabelsi **Philippe Langlais**

Année Universitaire / 2011-2012

Particularité de l'Inuktitut

- Langue très agglutinative.
  - ᖃᐃᓵᓕᓂᐊᖅᑰᙱᒃᑲᓗᐊᖅᐳᖅ (Martin *et al*, 2003)
  - **qaisaaliniaqquunngikkaluaqpuq**
    qai | saali | niaq | quu | nngik | kaluaq | puq
    nngit | galuaq
  - *"Actually he will probably not come early today"*

Corpus utilisé et système de traduction

- Hansards anglais et inuktitut de l'assemblée législative du Nunavut.
- 1er Avril 99 - 8 Novembre 07.
- 535000 phrases alignées.

Prétraitements

- 506162 phrases alignées.

	Occurrences	Mots disincts
Anglais	5603078	30749
Inuktitut	3153724	526344

Tableau 1 : Vocabulaire du Hansards

- **Vocabulaire Inuktitut 17 fois plus grand.**

Corpus utilisé et système de traduction

- Système de traduction statistique Moses avec fonctions standards. (Koehn *et al.*, 2007)
  http://www.statmt.org/moses/

- Entrainement : 470000 phrases alignées.

- Développement : 1000 phrases.

- Evaluation : 10000 phrases.

Résultats et significations

- 40 expériences relatives à 4 configurations.

Fichier Dev	Word Error Rate	Sentence Error Rate	BLEU
Dev1	44.77	**78.26**	**32.11**
Dev2	45.32	**79.86**	**30.90**
Dev3	45.21	**79.62**	**31.80**
Dev4	45.06	**80.51**	**31.00**
Moyenne	**45.10**	**79.60**	**31.43**

Tableau 2 : Résultats relatives à la traduction des textes parlementaires anglais-inuktitut.

- (Koehn, 2005) 110 systèmes ; 11 langues européennes; *Europarl v3*.

Résultats et significations

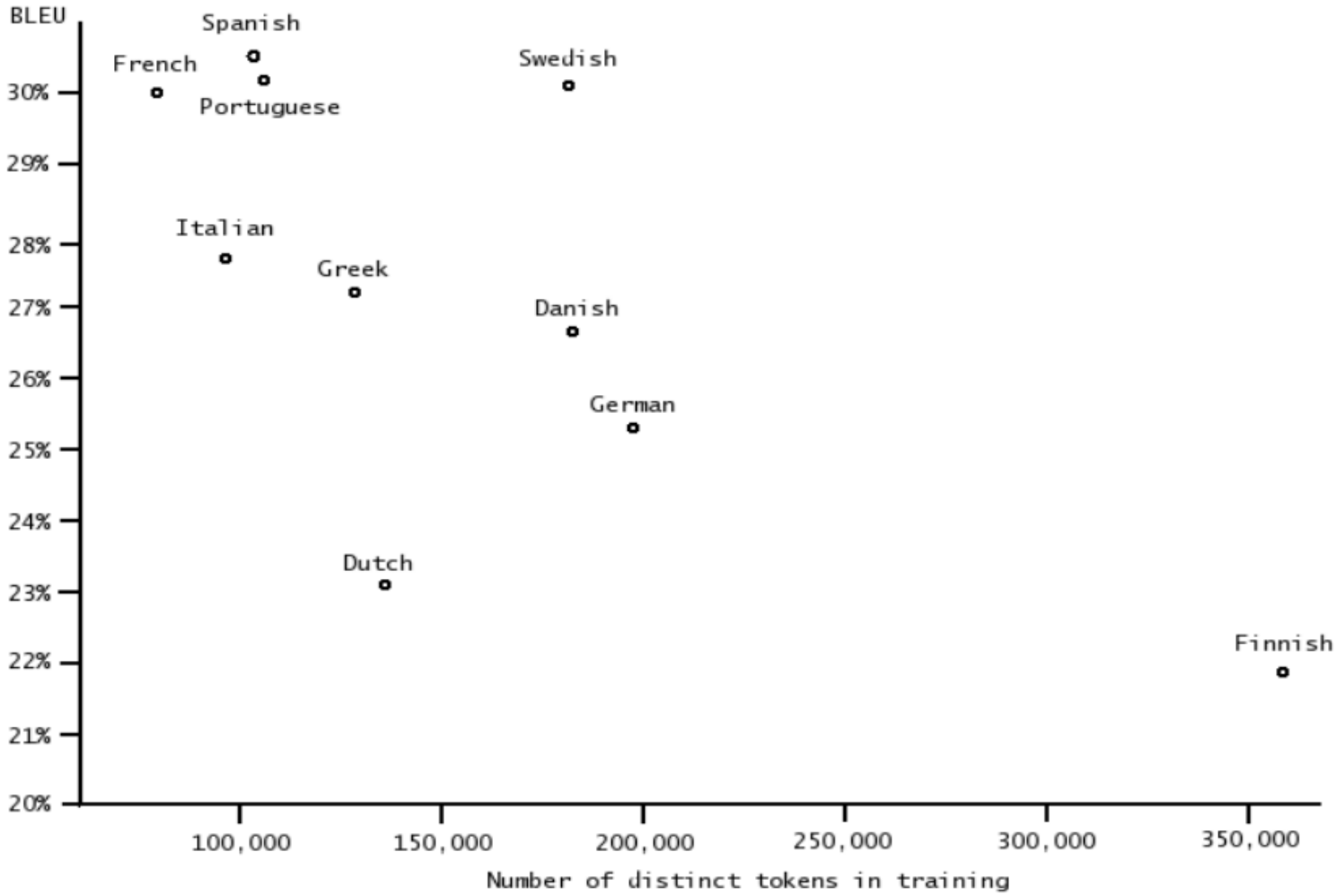

Figure 1 : Tailles des vocabulaires relatif à *Europarl v3* et scores Bleu dans la tâche traduction vers l'anglais. (Koehn, 2005)

Résultats et significations

- + difficile de traduire vers une langue morphologiquement riche.

Système de Traduction	Scores BLEU
Finnois-Anglais	21.8
Anglais-Finnois	13.0

Tableau 3 : Résultats des traductions relatives aux Finnois (Koehn, 2005)

- → Pas d'explication aux bons résultats anglais-inuktitut.

Résultats et significations

Langue	Occurrences d'évaluation	Occurrences Observées	Pourcentage
Anglais	113286	112949	99.67%
Inuktitut	70471	61699	87.50%

Tableau 4 :Proportion des occurrences d'évaluation observées à l'entrainement et au développement

- Analyse plus précise :
  - Mesurer la ressemblance entre les données d'évaluation et les données d'entrainement (données anglaises).
  - Comparer avec un corpus utilisé dans la littérature (*Europarl v6*).

Résultats et significations

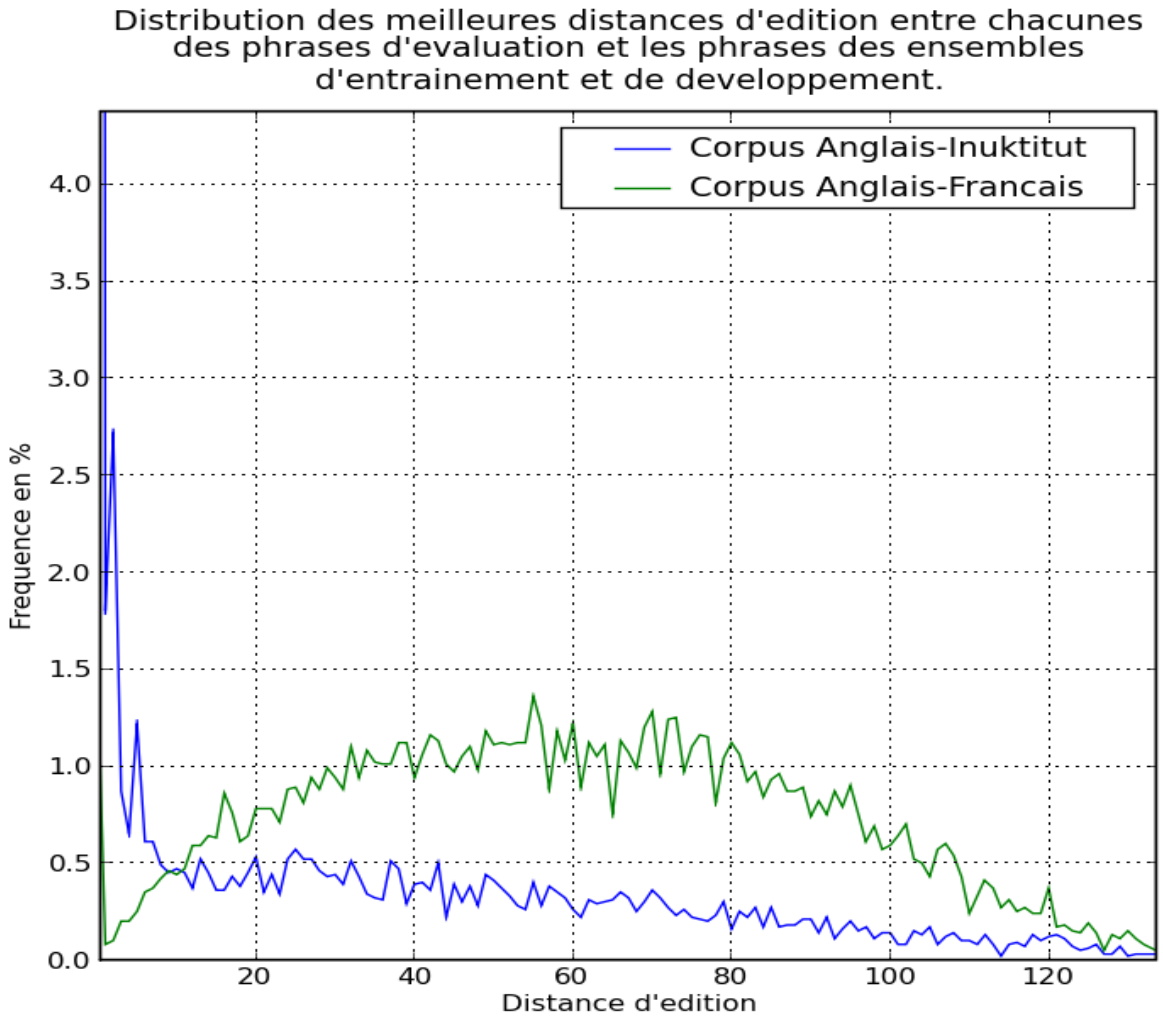

Figure 2 : Distributions des distances d'édition les plus proches des phrases d'évaluation

Résultats et significations

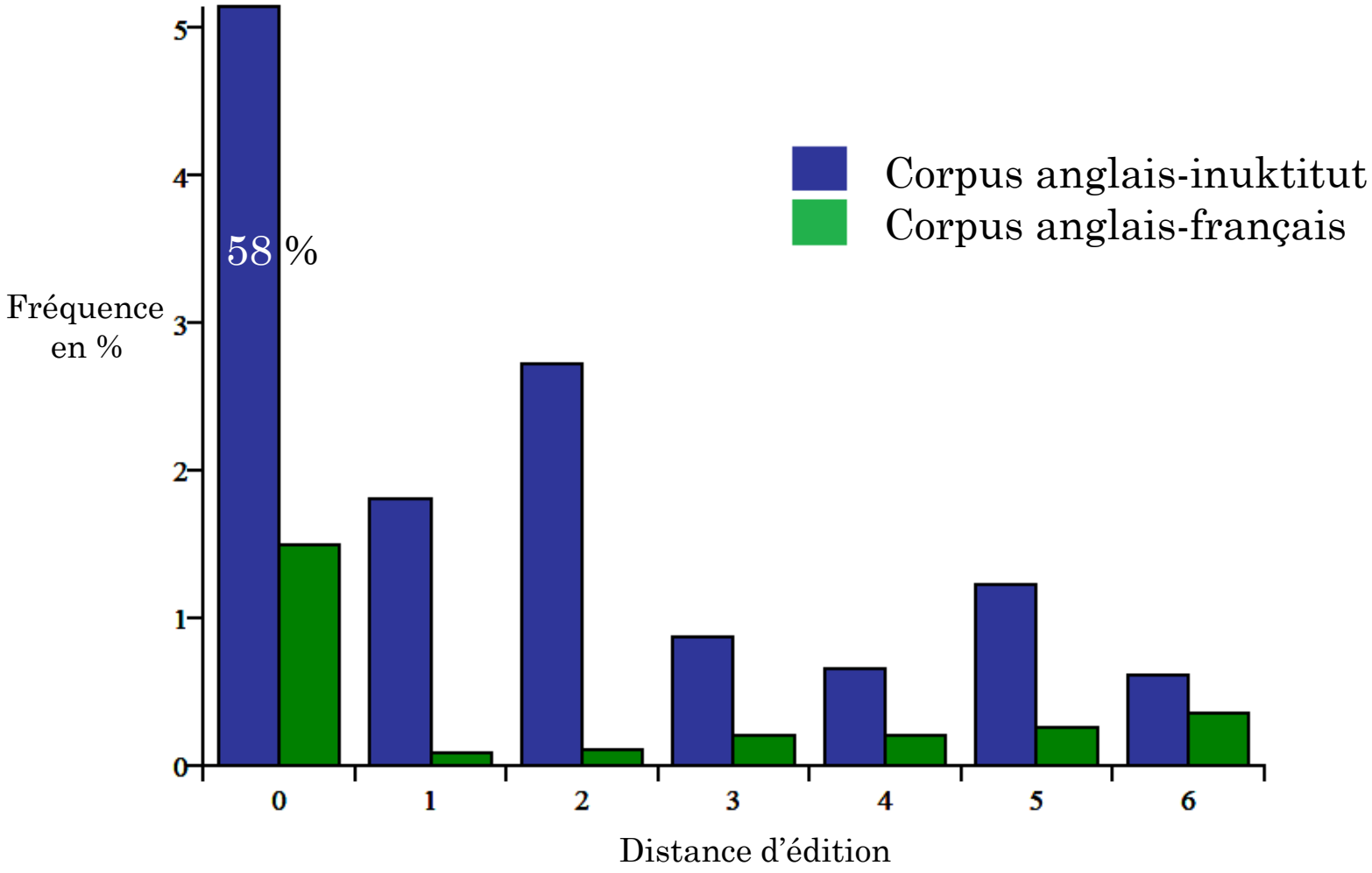

Figure 3 : Zoom sur l'histogramme des distances d'édition les plus proches des phrases d'évaluation

Données météorologiques

- Pas de références météorologiques pour l'inuktitut.
- 1000 phrases anglaises traduites. (*Warning bitext*) http://rali.iro.umontreal.ca/meteo/

Anglais	Traductions produites
" there were no tornados . "	*" pitaqangimmat tornados . "*
" this thunderstorm is moving eastsoutheastward at 50 km / h ."	*" tamanna thunderstorm doris eastsoutheastward 50-nik / km kamagijaksaqtaalaurmingmata . "*
"a tornado touchdown has been reported 12 km west of rimbey . "	*" tornado touchdown & & 12 km uangnaqpasianiglu rimbey ."*

Tableau 5 : Traductions de bulletins météorologiques

Données météorologiques

- Mots inconnus :
  - *cyxy, dakota, detroit, east-northeast, rainfall, rutherford, torrential, waterloo, wwcn15 …*

Vocab anglais	Vocab inuktitut	Mots inconnus
772	764	203 = 26.30% vocab anglais

Tableau 6 : Vocabulaires relatifs aux bulletins météorologiques

Travaux en cours

Source anglaise	Référence finnoise
they too now have a clear idea of the rights which they have to respect .	heilläkin on nyt selvä käsitys oikeuksista , joiden mukaisesti heidän pitää toimia .

Prétraitement	Post-traitement	Traduction générée	BLEU
Aucun	Aucun	ne on nyt selkeä käsitys oikeuksia , joita he ovat .	14.23
Stemming à 8 caractères	Prédiction morphologique	heidän on nyt selkeä ajatus oikeuksia , joita he ovat kunnioitetaan .	13.50
Stemming à l'aide de *Snowball*	Prédicion morphologique	he ovat myös nyt on selkeä käsitys oikeuksia , joita on kunnioitettava .	13.35
Segmentation	Aucun	myös heillä on nyt selvä käsitys oikeuksia , joita niiden on noudatettava .	14.68
Stemming à l'aide de la segmentation	Prédiction morphologique	ne on nyt selkeä käsitys oikeuksia , joita niiden on noudatettava .	14.80

Conclusions

- Système anglais-inuktitut
  - nécessite une adaptation à la traduction de bulletins.
  - besoin de références.

- Système anglais-finnois
  - travail en cours.
  - impact espéré pour l'inuktitut.

Référence

- J. Martin, H. Johnson, B. Farley, et A. Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. Dans *Proceedings of Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond,* HLT-NAACL 2003.
- Koehn, Philipp. *2005. Europarl*: A parallel corpus for statistical machine translation. *Dans Proceedings of Machine Translation Summit X*, *Association for Computational Linguistics*, *Phuket*, *Thailand*, 79–86.
- Ann Clifton et Anoop Sarkar. 2010. Unsupervised Morphological Segmentation for Statistical Translation.
- Alexandre Patry et Philippe Langlais. 2010. Intégration du contexte en traduction à l'aide d'un perceptron à plusieurs couches.
- William A. Gale et Kenneth Ward Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–103.
- Moore R. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. Dans *Machine Translation: From Research to Real Users. Actes* de *5th Conference of the Association for Machine Translation in the Americas, Tiburon, California, Springer-Verlag, Heidelberg*, *Germany*, 135-244.