



# Évolution du prototype de traducteur d'avertissements météo

Fabrizio Gotti 

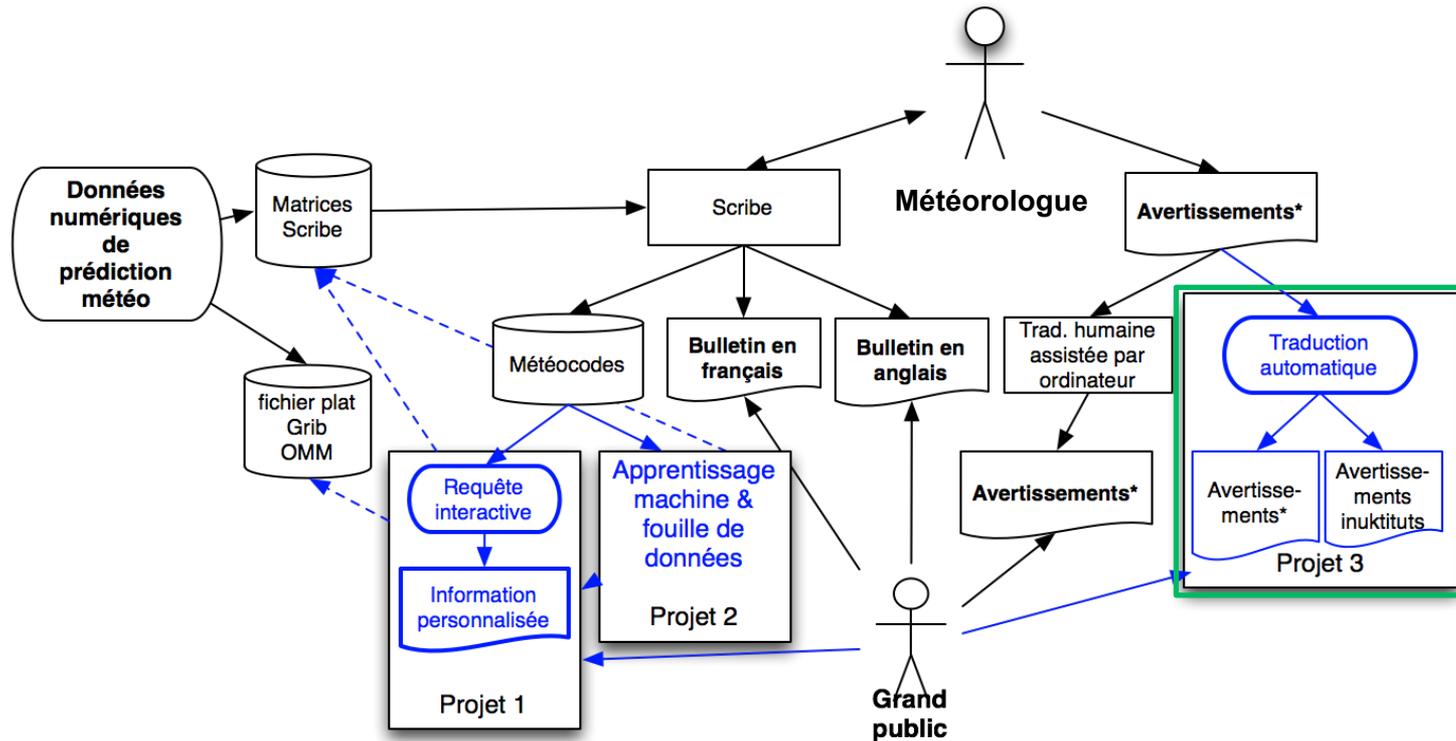
Philippe Langlais

Guy Lapalme

24 février 2012



# Traduction automatique des avertissements météo



Légende:



En noir : système actuel. En bleu : développements proposés



**WATT**

Warning-Avertissement Translation-Traduction

[In English](#)



Ce prototype a été développé afin de traduire les [avertissements météo publics](#) publiés par Environnement Canada.

Saisissez du texte dans la boîte de texte à gauche, puis cliquez sur Traduire, ou bien cliquez pour obtenir du texte d'exemple en [français](#) ou en [anglais](#)

### Texte source

Sunshine Coast  
4h33 HNP le vendredi 17 février 2012  
un avertissement de vent pour  
Sunshine Coast est émis

Vents du sud-est de 60 à 80 km/h se levant tôt cet après-midi.

Ceci est un avertissement indiquant que des vents potentiellement destructeurs soufflent ou souffleront dans ces secteurs. Veuillez Surveiller les conditions météo..Ainsi que les bulletins météorologiques et leurs mises à jour.

Une dépression qui s'approche continuera de se creuser à mesure qu'elle se déplacera pour se trouver sur le nord de l'île de Vancouver ce soir. À l'avant de la dépression, des vents du sud-est de 60 à 80 km/h se lèveront sur les secteurs côtiers de la Sunshine Coast et de l'île de Vancouver Est. Ces vents s'atténueront après minuit à mesure que la dépression envahira l'intérieur et faiblira.

### Traduction

Sunshine Coast, 4:33 AM PST Friday 17 february 2012  
Wind warning for Sunshine Coast issued.

Winds southeasterly 60 to 80 km/h developing early this afternoon.

This is a warning that potentially damaging winds are expected or occurring in these regions. Monitor weather conditions, listen for updated statements.

An approaching low pressure system will continue to deepen as it moves to lie over northern Vancouver Island this evening. Ahead of the low, southeast winds of 60 to 80 km/h will develop over coastal sections of the Sunshine Coast and East Vancouver Island. These winds will abate after midnight as the low moves inland and weakens.

Direction de traduction :

Partie du projet de [Diffusion multiformat d'informations environnementales](#).

© RALI 2011, Laboratoire de recherche appliquée en linguistique informatique

<http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html>

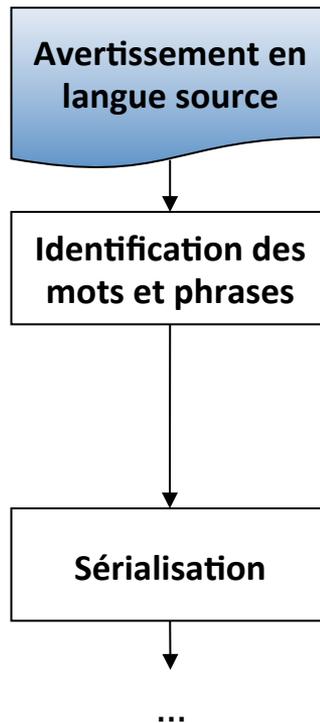


# Caractéristiques de WATT

- Développement depuis 2004
- Moteur de traduction statistique MOSES
  - Anglais ↔ français
  - Entraîné sur un corpus bilingue de bulletins et avertissements
- Disponible
  - Sur Linux (ligne de commande Python)
  - Sur le Web (interface graphique)
  - À l'écoute des fichiers déposés par ftp

# Fonctionnement de WATT (1)

## Discussion MTCN



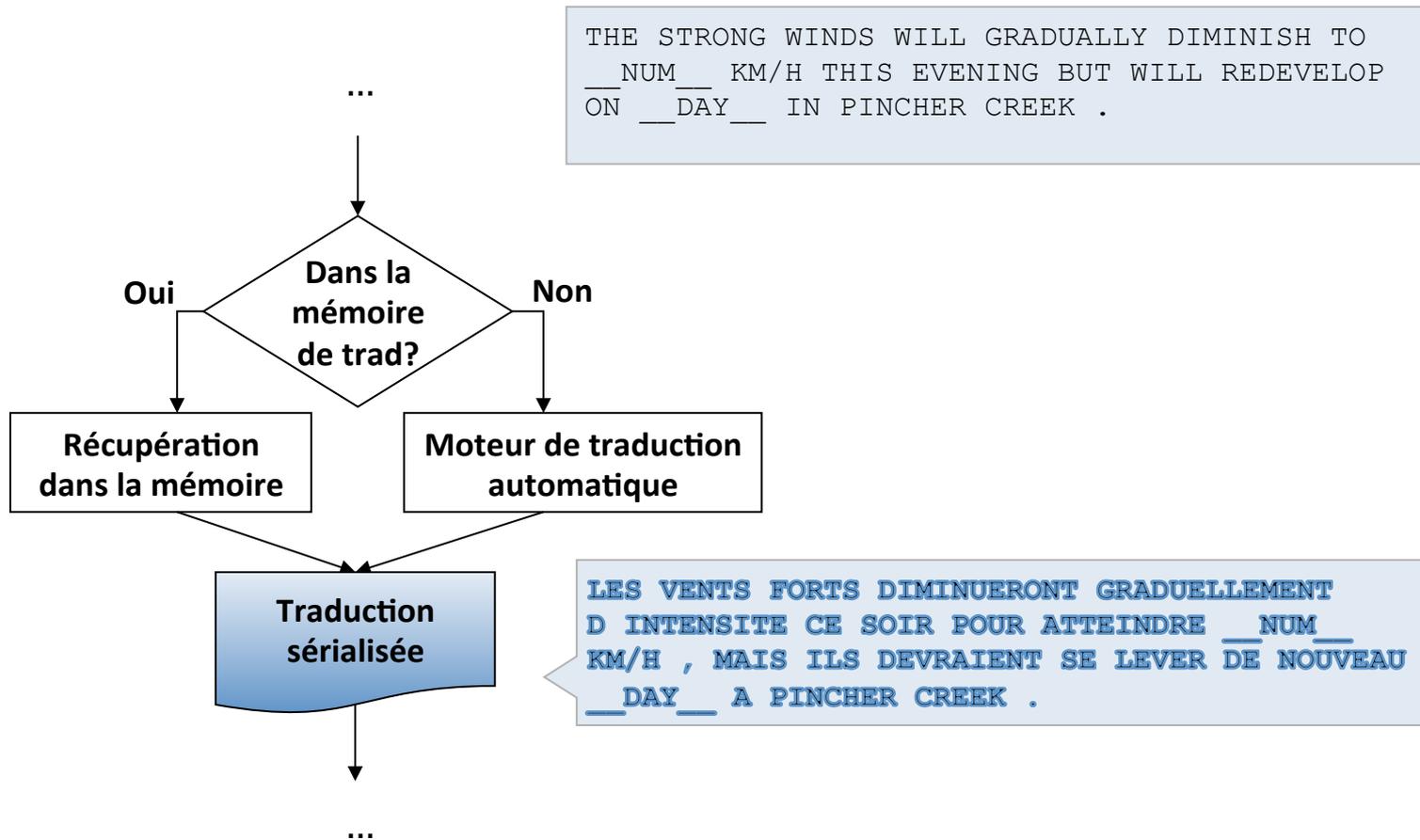
>>1>>

STRONG WESTERLY WINDS WITH GUSTS UP TO 100 KM/H ARE FORECAST TO DEVELOP IN THE PINCHER CREEK REGION THIS MORNING AND THEN SPREAD EASTWARD THROUGHOUT THE DAY. THE STRONG WINDS WILL GRADUALLY DIMINISH TO 30 KM/H THIS EVENING BUT WILL REDEVELOP ON WEDNESDAY IN PINCHER CREEK.

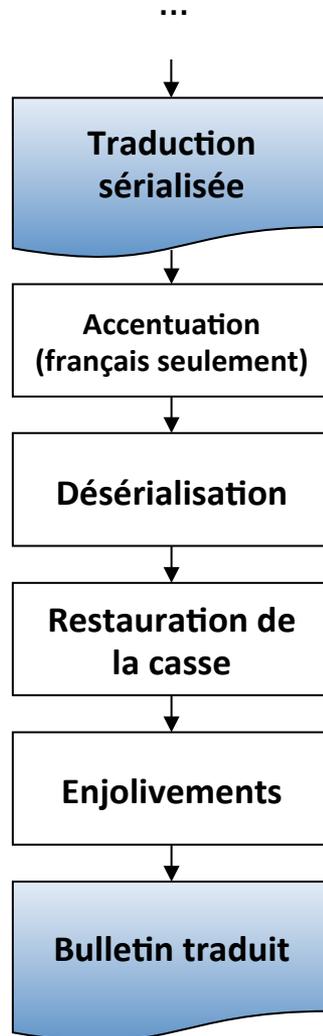
THE STRONG WINDS WILL GRADUALLY DIMINISH TO 30 KM/H THIS EVENING BUT WILL REDEVELOP ON WEDNESDAY IN PINCHER CREEK .

THE STRONG WINDS WILL GRADUALLY DIMINISH TO NUM KM/H THIS EVENING BUT WILL REDEVELOP ON DAY IN PINCHER CREEK .

# Fonctionnement de WATT (2)



# Fonctionnement de WATT (3)



LES VENTS FORTS DIMINUERONT GRADUELLEMENT D INTENSITE CE SOIR POUR ATTEINDRE NUM KM/H , MAIS ILS DEVRAIENT SE LEVER DE NOUVEAU DAY A PINCHER CREEK .

LES VENTS FORTS DIMINUERONT GRADUELLEMENT D **INTENSITÉ** CE SOIR POUR ATTEINDRE NUM KM/H , MAIS ILS DEVRAIENT SE LEVER DE NOUVEAU DAY **À** PINCHER CREEK .

LES VENTS FORTS DIMINUERONT GRADUELLEMENT D INTENSITÉ CE SOIR POUR ATTEINDRE **30** KM/H , MAIS ILS DEVRAIENT SE LEVER DE NOUVEAU **MERCREDI** À PINCHER CREEK .

**L**es vents forts diminueront graduellement d intensité ce soir pour atteindre 30 km/h , mais ils devraient se lever de nouveau mercredi à **Pincher Creek** .

Les vents forts diminueront graduellement d'intensité ce soir pour atteindre 30 km/h, mais ils devraient se lever de nouveau mercredi à Pincher Creek.

# Évaluations de 2010

- Deux évaluations (BT & RALI) découvraient des améliorations potentielles à moyen terme
  - Tournures particulières
  - Accentuation, casse (p.ex. celle des *noms de lieux*)
  - Unités de mesure
  - Améliorations du moteur de traduction
- La qualité du matériel source est cruciale
  - Mode de saisie des bulletins (Guy Lapalme)

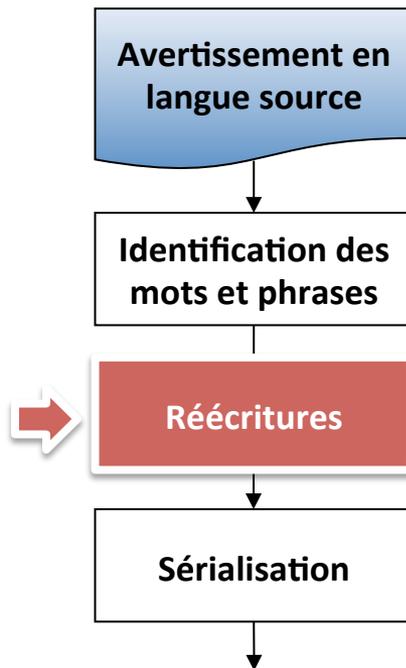
# Plus de données...

- Efforts d'acquisition de plus de données

Tableau 1 – Données source (toutes les données en milliers)

Source	Paires de doc.	Paires de phrases	Paires différentes	Anglais		Français	
				Mots	Mots diff.	Mots	Mots diff.
Prévisions	90	4187	349	30296	6.9	37284	8.0
Avertissements 2000-2004	30	105	70	1733	6.0	2067	8.2
Avertissements 2005-2009	51	235	130	4160	7.3	4871	9.8
Avertissements 2009-2011	35	332	87	5177	7.1	6435	8.8
Tous avertissements	116	672	281	11 070	11.1	13 372	14.8
Bitexte complet	<b>205</b>	<b>4 859</b>	<b>631</b>	<b>41 366</b>	<b>15.2</b>	<b>50 656</b>	<b>19.4</b>

# Amélioration au prétraitement



- > 50 règles de réécriture

“P.E.I.” → “PRINCE EDWARD ISLAND”

“JAN 1ST” → “1 JANUARY”

“ENTRE 3 ET 4 H” → “ENTRE 3 H ET 4 H”

“QUARTER SIZED HAIL” → “24 MM HAIL”

MARBLE SIZED, MARBEL SIZED

GRAPE SIZED

DIME SIZE HAIL

PENNY SIZE

SIZE OF NICKELS

THE SIZE OF QUARTERS

THE SIZE OF LOONIES, LOONIE SIZED

TOONIE SIZED, \*TWOONIE SIZED, \*TWOONIE SIZED

WALNUT SIZED

PING PONG SIZED

GOLF BALL, GOLFBALL-SIZED

EGG SIZED

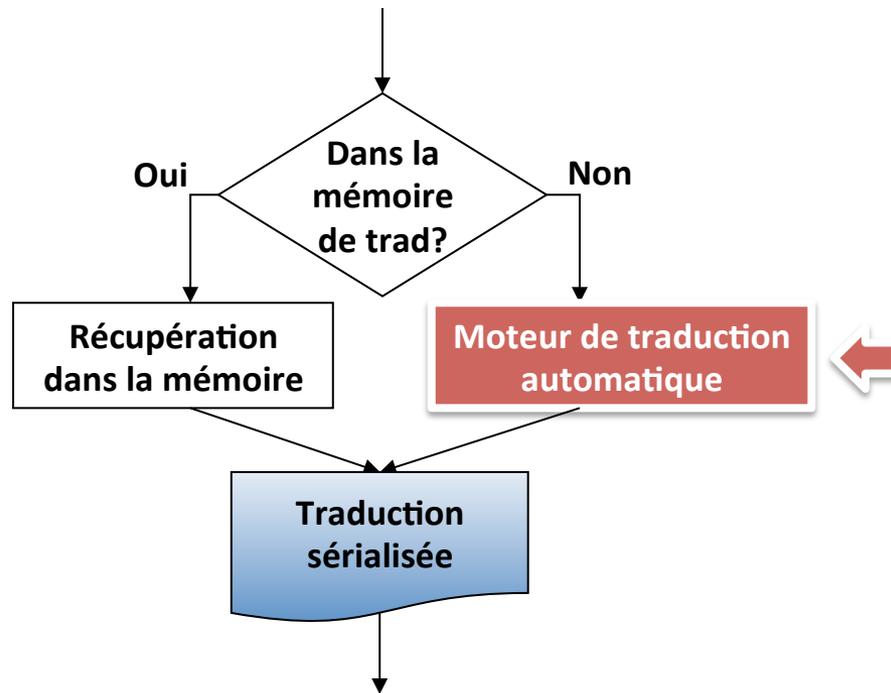
TENNIS-BALL SIZED, TENNIS BALL SIZED

BASEBALL SIZED

SIZE OF SOFTBALLS

...

# Améliorer Moses



# Ré-entraînement de Moses

- Constitution de trois sous-corpus
  - *Train* pour l'entraînement **moins** hiver 2010 et été/automne 2011
  - *Tune* et *test* **seulement** sur 400 bulletins difficiles de hiver 2010 et été/automne 2011

Tableau 2 – Partition entraînement (toutes les données en milliers)

Corpus	Paires de phrases	Anglais			Français		
		Phrases diff.	Mots	Mots uniques	Phrases diff.	Mots	Mots uniques
Train	4 761	523	39 906	15	559	48 847	19
Tune	3.8	1.7	59.7	1.8	1.7	73.8	2.1
Test	3.7	1.8	57.8	1.8	1.8	71.4	2.1

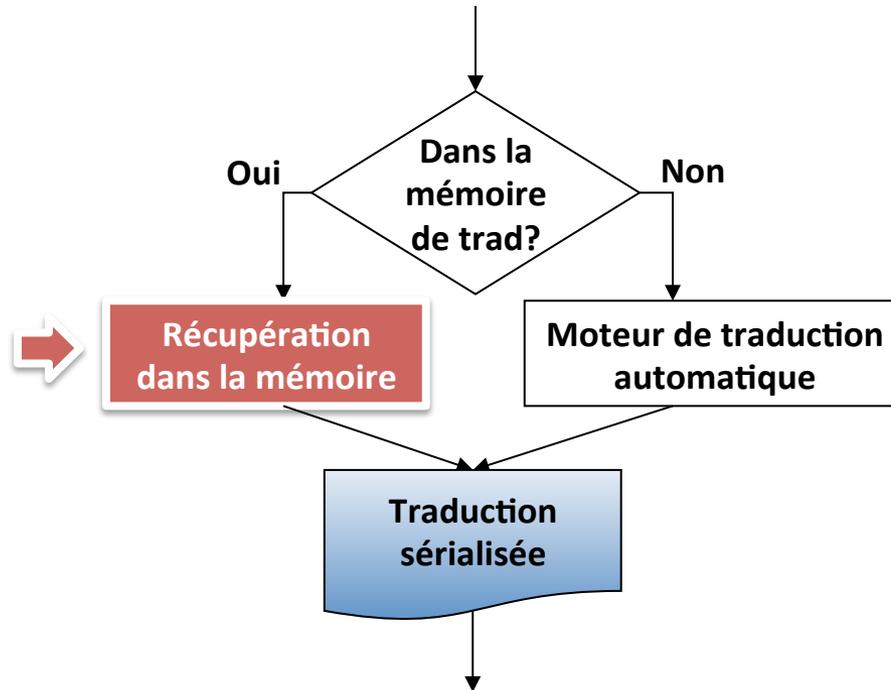
# Tests de 9 configurations différentes

- 3 stratégies de réordonnement des mots
  - Aucune
  - Basée sur la distance
  - ★ - Basée sur des opérations lexicalisées (msd)
- 3 filtres des associations du modèle de Moses

BALAIE ↔ CHIBOUGAMAU TOWARD LAKE ERIE WILL SWEEP

  - différence de longueur doit  $\leq 4$
  - ★ - différence de longueur doit  $\leq 5$
  - aucun filtre
- On mesure les résultats avec BLEU (%) sur le test ( $n = 5$  expérimentations)

# Intégrer une mémoire de traduction



# Deux versions de la mémoire

Français	Anglais	Fréquence
MAXIMUM DE _NUM_ .	HIGH _NUM_ .	190 925
CECI EST UN AVERTISSEMENT INDIQUANT QUE DES ORAGES VIOLENTS SONT SUR LE POINT D AFFECTER OU AFFECTENT DEJA CES REGIONS .	THIS IS A WARNING THAT SEVERE THUNDERSTORMS ARE IMMINENT OR OCCURRING IN THESE REGIONS .	13 525
CETTE ALERTE EST EN VIGUEUR DE _TIME_ A _TIME_ HAE .	THIS WARNING IS IN EFFECT FROM _TIME_ TO _TIME_ EDT .	2826
...		
A COLD FRONT MOVING THROUGH THIS UNSTABLE AIR WILL CAUSE POTENTIALLY SEVERE THUNDERSTORMS OVER SEVERAL REGIONS .	LE PASSAGE D UN FRONT FROID DANS CET AIR INSTABLE DONNE DES ORAGES POTENTIELLEMENT VIOLENTS SUR PLUSIEURS REGIONS .	5
A BAND OF CLOUD PRODUCING SNOW AND ICE CRYSTALS WILL MOVE ACROSS NORTHERN MANITOBA THIS MORNING .	UNE BANDE DE NUAGES PRODUISANT DE LA NEIGE ET DES CRISTAUX DE GLACE SE DEPLACERA SUR LE NORD DU MANITOBA CE MATIN .	1
_NUM_ CM AU COUR DE LA PROCHAINE JOURNEE	_NUM_ CM MORE ARE LIKELY OVER THE NEXT DAY	1

95 k paires  
*fréquence > 5*

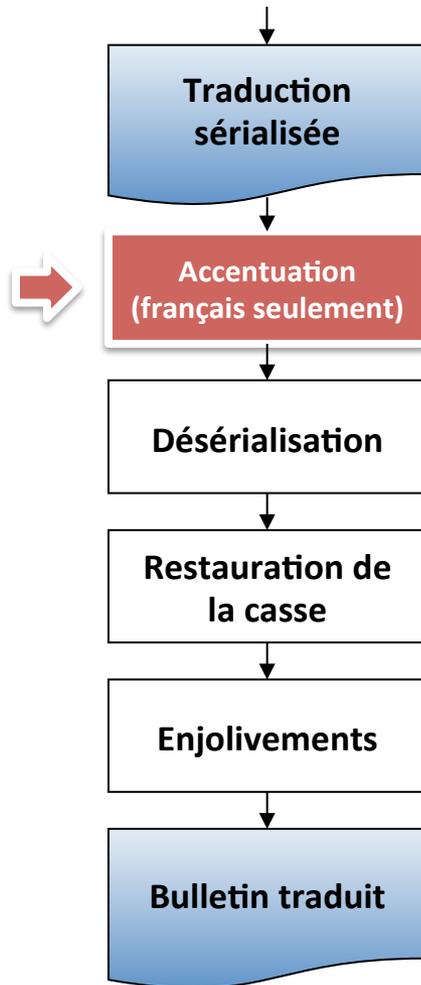
290 k paires  
*toutes les paires*

# Mémoire – résultats

Langue source	Système	Score BLEU (%)	WER (%)	SER (%)	% phrases en mémoire
Anglais	SMT (baseline)	79.6	14.7	58.1	n/a
	SMT+MEM	80.0	13.8	55.7	45.8
	SMT+MEM <sub>small</sub>	80.0	13.8	55.9	45.2
Français	SMT (baseline)	78.0	13.9	54.2	n/a
	SMT+MEM	78.4	13.3	52.6	44.6
	SMT+MEM <sub>small</sub>	78.4	13.3	52.7	44.1

- Le système SMT+MEM améliore les résultats et semble raisonnable

# Améliorer le post-traitement : accents

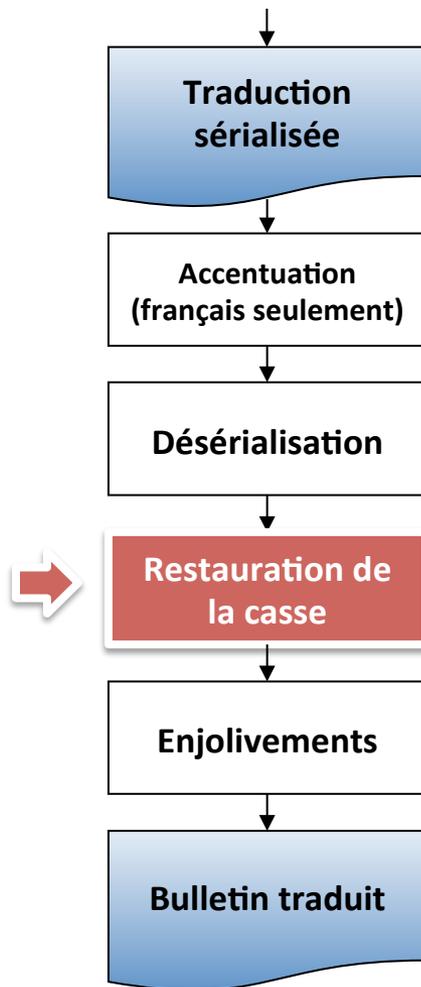


- Restauration des accents à l'aide de listes de mots accentués, guidée par un modèle statistique

il a neige a Montreal → il a neigé à Montréal

- Ajout d'un lexique à plus large couverture (un peu plus lent...)

# Améliorer le post-traitement



- Restauration de la casse   partir de *listes de noms de lieux (canadiens)*

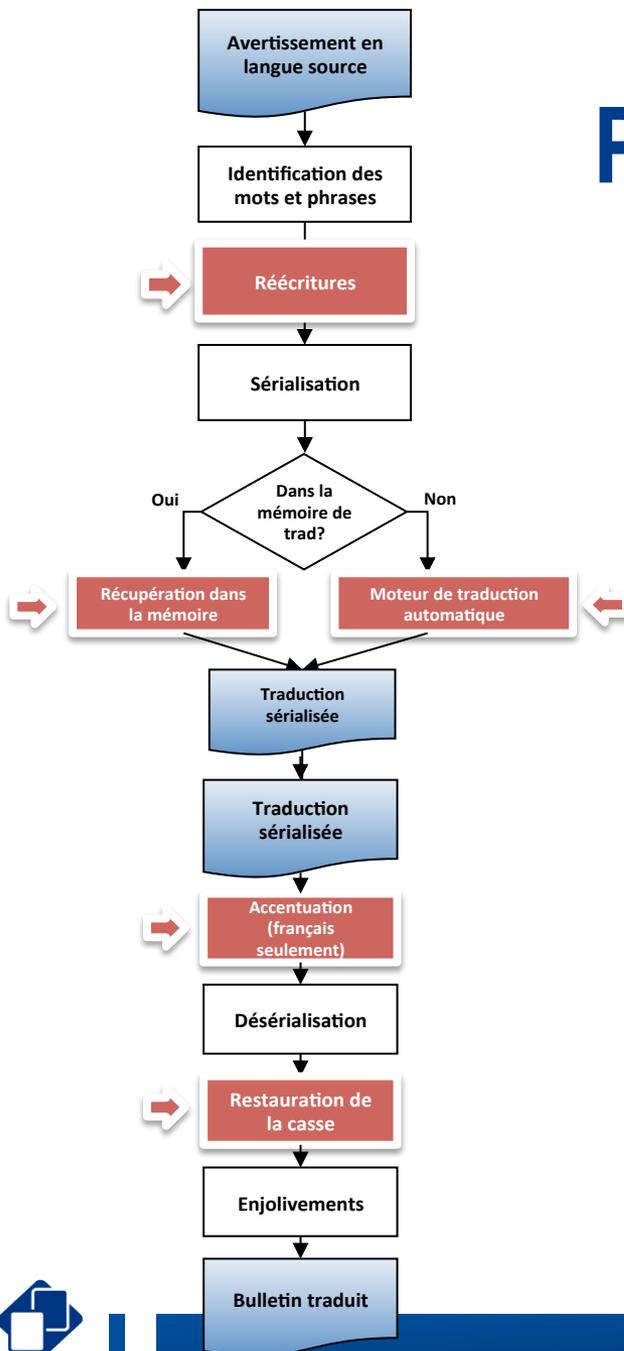
AN AREA NEAR METSO A CHOOT INDIAN RESERVE 23 →  
an area near Metso A Choot Indian Reserve 23

- DBpedia (900 fr, 4500 en)
- Avertissements publics d’EC (1500)
- GeoBase (16 000)
- M t ocode (500)
- Avertissements maritimes d’EC (400)
- Additions manuelles (UV, MTCN,...)

- Il reste des difficult s : HIGH LEVEL

Pop : 3641

# Portrait complet



- Améliorations à différents niveaux du pipeline de traduction
- Tourne sur un PC de bureau
- 5 s par phrase en moyenne
- Conception facile d'entretien (modifiable aisément)

# Conclusion

- La majorité des améliorations potentielles identifiées ont été apportées et permettent une performance de WATT très intéressante

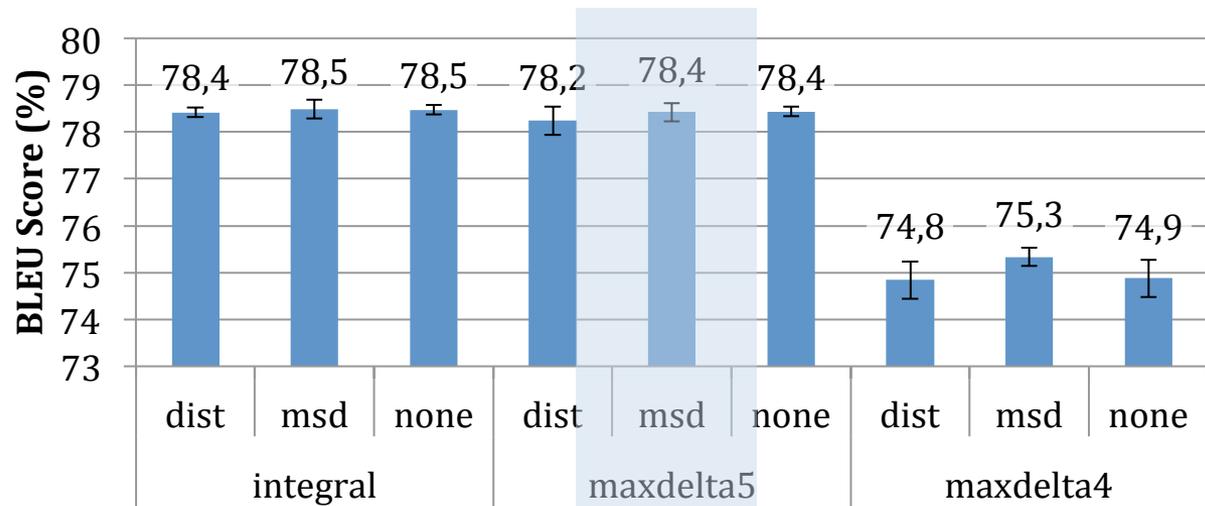
→ Rapport sur ceci transmis à EC le 6 janvier 2012

- Évaluation humaine
- Intégration chez EC

*Merci*

# Tests de 9 configurations – résultats

→ français



→ anglais

