Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

UPDATING WATT'S TRANSLATION MEMORY
PART 1: FINDINGS ON THE CAP DATA

Fabrizio Gotti – gottif@iro.umontreal.ca

Guy Lapalme – lapalme@iro.umontreal.ca

15 FEBRUARY 2016

RALI

Recherche Appliquée en
Linguistique Informatique

rali.iro.umontreal.ca

Université de Montréal

1 Table of contents

2 Introduction

On 6 January 2016, EC sent to RALI an archive of weather alerts published on Datamart from 2013 to January 2016, in Common Alerting Protocol (CAP) format[1]. RALI wanted to update its WATT translation engine (summarily described below) with the discussion text available in those alerts.

Prior to reporting how this new source of texts impacts WATT, we wanted to describe the data received and identify potential issues with it. This report is the result of this analysis.

Our discussion includes recommendations pertaining to this data as well as to the process that generated it.

2.1 Context of the study: WATT translation engine

Since 2008, RALI has been working on the dissemination of environmental information[2]. This project led to the creation of WATT[3], a prototype dedicated to the translation of public weather warnings issued by Environment Canada.

This prototype has been documented in an article (Gotti, Lapalme, & Langlais, 2014) entitled *Designing a Machine Translation System for Canadian Weather Warnings: a Case Study*, published in the journal *Natural Language Engineering* in 2014[4].

WATT combines a statistical machine translation engine (SMT) and a translation memory (TM). When a source sentence is submitted to the engine, it is first looked up in the TM. If it is present, the corresponding translation is retrieved and output. If the sentence is not found, the sentence is sent over to the SMT. The TM contains high-quality sentence pairs, mined from human translations.

2.2 Data soundness is vital to reliable machine translation

We have spent substantial effort here analyzing the textual data received in the archive and pointing out potential issues. This is necessary because the soundness of this data is critical to the development of a reliable machine translation system. In the case of WATT, the original text and its translation form the basis of the memory used to look up past translations, as explained above. Moreover, this bitext is used to train the statistical machine translation engine. Systematic errors in this data are bound to negatively affect the resulting system.

---

[1] http://docs.oasis-open.org/emergency/cap/v1.2/CAP-v1.2-os.html
[2] http://rali.iro.umontreal.ca/rali/?q=en/EnvironmentalInfo
[3] http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html
[4] http://rali.iro.umontreal.ca/rali/sites/default/files/publis/WATT-NLE-Final.pdf

3 Datamart archive

3.1 Archive content

The archive we received was a tarball containing 1057 smaller archives, one for each day. Two of these smaller archives were corrupted:

- 2013010400_tar.gz
- 2013010600_tar.gz

We decompressed everything and found 126,926 CAP files, in XML format. Two of those files were not well-formed (invalid XML structure):

- T_WWCN17_C_CWHX_201301061952_3FDE269D.cap
- T_WWCN11_C_CWTO_201301041457_193735EE.cap

R1. We recommend double-checking the archive creation process on Datamart, for CAP alerts, so as to make sure that all the required data is properly backed up.

3.2 CAP alerts

We found 126,926 usable CAP alerts. For each alert, we relied on the file name to determine the following information:

- Alert type (WW, WU, etc.)
- Emitting station
- Date

We spot-checked a few alerts to verify that the file name scheme does indeed reflect the content of the CAP file. The results were consistent.

Table 1 shows the distribution of CAP alerts found in the archive, by bulletin types and emitting stations.

Since we are only interested in the weather warnings containing text, we focus here on alerts with a text description. The statistics after filtering for this content are shown in Table 2. We further restrict the warnings we are interested in to WW and WO warnings, since those are the warnings WATT has been trained on since the very beginning of the project. According to our documentation, WO warnings are "pre-warnings and preparation information" and WW are proper weather warnings.

These WO and WW warnings account for 71,699 of the warnings received. Figure 1 shows the distribution by month for these warnings, from 2013 to 2016.

Table 1 – Number of CAP alerts with and without text content, by bulletin type and emitting station.

	Station							
Bulletin type	**HX**	**NT**	**QX**	**TO**	**UL**	**VR**	**WG**	**Total**
WE						6		6
WF	11			255	48		499	813
WO	8 028	598	150	6 796	2 500	3 493	3 022	24 587
WS	257	8		234	131	92	211	933
WT	39							39
WU	544	8		3 936	1 749	497	10 697	17 431
WW	21 188	8 373		15 380	13 731	8 578	15 867	83 117
Total	**30 067**	**8 987**	**150**	**26 601**	**18 159**	**12 666**	**30 296**	**126 926**

Table 2 – Number of CAP alerts with content (text description), by bulletin type and emitting station.

	Station							
Bulletin type	**HX**	**NT**	**QX**	**TO**	**UL**	**VR**	**WG**	**Total**
WE								
WF	5			132	29		294	460
WO	5 725	393	71	4 237	1 794	2 643	2 311	17 174
WS	234	6		190	113	73	177	793
WT	19							19
WU	351	4		2 404	1 159	323	6 750	10 991
WW	13 687	5 722		10 133	8 999	5 926	10 058	54 525
Total	**20 021**	**6 125**	**71**	**17 096**	**12 094**	**8 965**	**19 590**	**83 962**

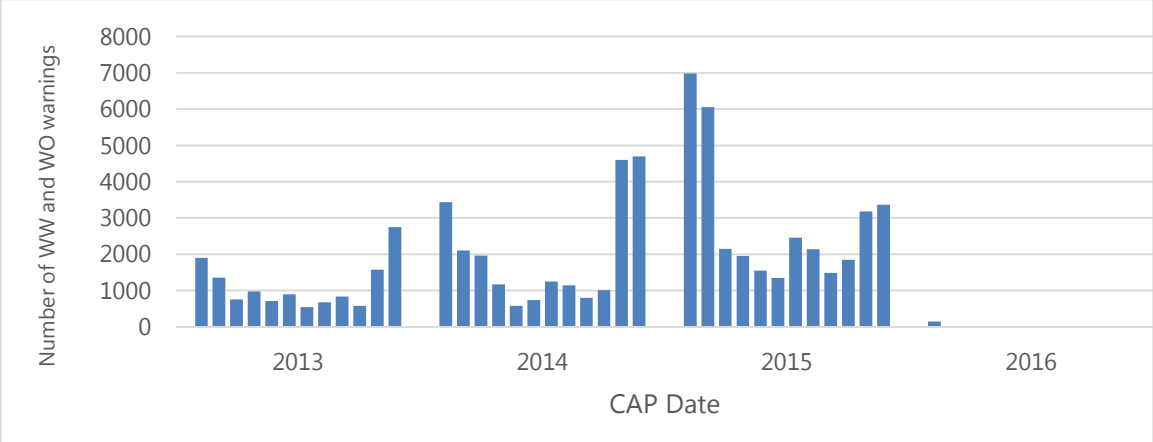

Figure 1 – Number of WW and WO warnings with text (description), by date. Each bar corresponds to a month.

4 Corpus preparation

4.1 Text preparation steps

We extract the descriptions from the CAP warnings in order to produce a corpus of usable bilingual texts (bitexts). The steps to achieve this are:

1. the French and English descriptions are extracted from the CAPs
2. the whitespace is removed and normalized
3. the texts are tokenized (segmented into sentences, then words)
4. the texts are serialized, so that (for instance) numbers are all folded into a single token (__NUM__), in order to maximize recall when the translation memory is eventually produced.[5]
5. the French and English sentences are aligned, so that we obtain pairs of sentences that are a translation of one another.

The text alignment is performed with our tool `yasa`, available online[6].

Figure 3 illustrates the text extraction pipeline starting from the CAP files. The tokenized version of the text follows WATT's internal format, i.e. all uppercase letters, with some characters removed (e.g. apostrophes). This internal format is inspired from the format of MTCN bulletins (see Figure 2 below for an excerpt), a non-XML format.

```
>>1>>
<DISCUSSION>.

>>2>>
PLEASE REFER TO THE LATEST PUBLIC FORECASTS FOR FURTHER DETAILS.

>>3>>
SCATTERED THUNDERSTORMS WITH LOCALIZED DAMAGING WINDS AND LARGE HAIL ARE MOVING
EASTWARD THROUGH OR TOWARD PORTIONS OF THE ABOVE REGIONS. THERE IS THE RISK OF A
TORNADO WITH SOME OF THESE STORMS.

>>4>>
THIS IS A WARNING THAT SEVERE THUNDERSTORMS ARE IMMINENT OR OCCURRING IN THESE
REGIONS. REMEMBER SOME SEVERE THUNDERSTORMS PRODUCE TORNADOES..LISTEN FOR UPDATED
WARNINGS. EMERGENCY MANAGEMENT ONTARIO RECOMMENDS TAKING COVER IMMEDIATELY WHEN
THREATENING WEATHER APPROACHES.

>>5>>
NOTE..A SUMMARY OF ALL WARNINGS AND WATCHES FOR SOUTHERN ONTARIO IS AVAILABLE IN THE
WWCN11 CWTO BULLETIN ISSUED IMMEDIATELY FOLLOWING THIS BULLETIN.

[...]
```

Figure 2 – Excerpt from a sample MTCN bulletin dated 11 July 2009, from RALI's archive.

[5] For more information on this step, consult (Gotti, Lapalme, & Langlais, 2014), section 4.1.

[6] http://rali.iro.umontreal.ca/rali/?q=en/yasa

Step 1. CAP file description (XML)	
<description>Blizzard conditions with poor visibility in snow and blowing snow are expected or occurring. Blizzard conditions over the Churchill and York regions are expected to continue tonight and into Friday morning. Very strong northwest winds of 50 km/h with gusts to 70 km/h have developed early this afternoon bringing blizzard conditions to the Churchill and York regions. Weather conditions will improve Friday morning.</description>	<description>Il y a ou il y aura du blizzard avec une visibilité mauvaise sous la neige et dans la poudrerie. Blizzard sur les régions de Churchill et de York devraient continuer ce soir, cette nuit et vendredi matin. Des vents très forts du nord-ouest de 50 km/h avec rafales à 70 km/h se sont levés tôt cet après-midi apportant du blizzard sur les régions de Churchill et de York. La situation s'améliorera vendredi matin.</description>
Steps 2. and 3. Text extraction and whitespace normalization	
Blizzard conditions with poor visibility in snow and blowing snow are expected or occurring. Blizzard conditions over the Churchill and York regions are expected to continue tonight and into Friday morning. Very strong northwest winds of 50 km/h with gusts to 70 km/h have developed early this afternoon bringing blizzard conditions to the Churchill and York regions. Weather conditions will improve Friday morning.	Il y a ou il y aura du blizzard avec une visibilité mauvaise sous la neige et dans la poudrerie. Blizzard sur les régions de Churchill et de York devraient continuer ce soir, cette nuit et vendredi matin. Des vents très forts du nord-ouest de 50 km/h avec rafales à 70 km/h se sont levés tôt cet après-midi apportant du blizzard sur les régions de Churchill et de York. La situation s'améliorera vendredi matin.
Step 4. Tokenization and serialization	
BLIZZARD CONDITIONS WITH POOR VISIBILITY IN SNOW AND BLOWING SNOW ARE EXPECTED OR OCCURRING . BLIZZARD CONDITIONS OVER THE CHURCHILL AND YORK REGIONS ARE EXPECTED TO CONTINUE TONIGHT AND INTO **__DAY__** MORNING . VERY STRONG NORTHWEST WINDS OF **__NUM__** KM / H WITH GUSTS TO **__NUM__** KM / H HAVE DEVELOPED EARLY THIS AFTERNOON BRINGING BLIZZARD CONDITIONS TO THE CHURCHILL AND YORK REGIONS . WEATHER CONDITIONS WILL IMPROVE __DAY__ MORNING .	IL Y A OU IL Y AURA DU BLIZZARD AVEC UNE VISIBILITE MAUVAISE SOUS LA NEIGE ET DANS LA POUDRERIE . BLIZZARD SUR LES REGIONS DE CHURCHILL ET DE YORK DEVRAIENT CONTINUER CE SOIR , CETTE NUIT ET **__DAY__** MATIN . DES VENTS TRES FORTS DU NORD-OUEST DE **__NUM__** KM / H AVEC RAFALES A **__NUM__** KM / H SE SONT LEVES TOT CET APRES-MIDI APPORTANT DU BLIZZARD SUR LES REGIONS DE CHURCHILL ET DE YORK . LA SITUATION S AMELIORERA __DAY__ MATIN .
Step 5. Sentence alignment and resulting bitext	
BLIZZARD CONDITIONS WITH POOR VISIBILITY IN SNOW AND BLOWING SNOW ARE EXPECTED OR OCCURRING .	IL Y A OU IL Y AURA DU BLIZZARD AVEC UNE VISIBILITE MAUVAISE SOUS LA NEIGE ET DANS LA POUDRERIE .
BLIZZARD CONDITIONS OVER THE CHURCHILL AND YORK REGIONS ARE EXPECTED TO CONTINUE TONIGHT AND INTO __DAY__MORNING .	BLIZZARD SUR LES REGIONS DE CHURCHILL ET DE YORK DEVRAIENT CONTINUER CE SOIR , CETTE NUIT ET __DAY__ MATIN .
VERY STRONG NORTHWEST WINDS OF __NUM__ KM / H WITH GUSTS TO __NUM__ KM / H HAVE DEVELOPED EARLY THIS AFTERNOON BRINGING BLIZZARD CONDITIONS TO THE CHURCHILL AND YORK REGIONS .	DES VENTS TRES FORTS DU NORD-OUEST DE __NUM__ KM / H AVEC RAFALES A __NUM__ KM / H SE SONT LEVES TOT CET APRES-MIDI APPORTANT DU BLIZZARD SUR LES REGIONS DE CHURCHILL ET DE YORK .
WEATHER CONDITIONS WILL IMPROVE __DAY__ MORNING .	LA SITUATION S AMELIORERA __DAY__ MATIN .

Figure 3 – Example of text extraction pipeline for a single CAP file. Excerpt from file T_WWCN12_C_CWWG_201501082025_BA5AF4B7.cap, with English as the original.

4.2 Corpus statistics

When the process described in the previous section is carried out on every CAP file from WO and WW warnings (71,699 files), we obtain an aligned bitext in the format expected by the WATT translation pipeline (an excerpt is shown at the bottom of Figure 3).

Table 3 shows the statistics for the end result. The CAP files yield 375,225 aligned sentence pairs. The French corpus counts 7.3M tokens and the English one, 6M. This difference is expected. There is a lot of repetition in the corpus, which is also usual. On average, each English sentence is repeated 4.7 times, and each French sentence, 4.0 times. However, this average is misleading, since (in English) a few sentences are repeated more than 10,000 times (e.g. `ENVIRONMENT CANADA METEOROLOGISTS WILL UPDATE ALERTS AS REQUIRED .`), while about 40,000 sentences appear only once.

On average, an English bulletin counts 83 tokens; a French bulletin has 102 tokens per bulletin. In both case, this is about 5 sentences.

Table 3 – Bitext (aligned corpus) statistics. The text is serialized.

Language	Bulletins	Sents	Unique Sents	Tokens	Types	Avg. Toks per Sent	Avg. Toks per bulletin
English	71 699	375 225	79 774	5 971 531	6 963	16	83
French	71 699	375 225	94 637	7 319 508	8 531	20	102

Since these two corpora are sentence-aligned, they can be readily converted into a translation memory, where the basic unit is not a sentence, but a sentence pair. Each sentence pair consists of an English sentence and its corresponding translation (or vice versa), along with the frequency with which it is found in the aligned corpora described in Table 3. A short excerpt of the translation memory (TM) is shown in Table 4.

Table 4 – Random excerpt of the translation memory derived from the CAP bitext

Freq	English sentence	French sentence
13 847	ENVIRONMENT CANADA METEOROLOGISTS WILL UPDATE ALERTS AS REQUIRED .	LES METEOROLOGISTES D ENVIRONNEMENT CANADA METTRONT LES ALERTES A JOUR AU BESOIN .
29	CONDITIONS ARE NO LONGER EXPECTED TO REACH HEAT WARNING CRITERIA .	ON NE PREVOIT PLUS DE CONDITIONS JUSTIFIANT L EMISSION D UN AVERTISSEMENT DE CHALEUR .
1	WINTRY WEATHER EXPECTED OVER HIGHER TERRAIN ON __DAY__ INTO EARLY __DAY__ .	ON PREVOIT DU TEMPS HIVERNAL **SUR EN** SECTEURS MONTAGNEUX __DAY__ ET TO T __DAY__ .
1	WINTRY WEATHER EXPECTED OVER HIGHER TERRAIN ON __DAY__ INTO EARLY __DAY__ .	ON PREVOIT DU TEMPS HIVERNAL EN TERRAIN MONTAGNEUX __DAY__ ET TOT __D AY__ .

We also summarize additional statistics for the TM in Table 5. There are 100K (different) sentence pairs in the TM. Typically, only those appearing more than *n* times are added to WATT's TM, where *n* is found empirically. This is done to increase the quality of the TM, since more frequent sentence pairs generally indicate a better quality, but the method significantly reduces the number of usable pairs. If we set *n* = 1, we lose two thirds of the original TM, for instance. The second to last row in Table 4 shows how a sentence pair with frequency 1 contains an error (**SUR EN** TERRAIN MONTAGNEUX).

Table 5 – Statistics for the translation memory

Nb of sentence pairs	99 532
Nb of sentence pairs with freq > 1	36 663
Nb of sentence pairs with freq > 3	9 786
Avg nb of FR sentences for each EN sentence	1.25
Avg nb of EN sentences for each FR sentence	1.05

An interesting statistic also shown in Table 5 is the average number of different French sentences associated with the same English sentence (and vice-versa). As illustrated in the last two rows of Table 4, a single sentence (WINTRY WEATHER EXPECTED OVER HIGHER TERRAIN ON __DAY__ INTO EARLY __DAY__ .) can have more than a single equivalent in the other language. On average, a single French sentence has 1.05 English equivalents and a single English sentence can have 1.25 French equivalents.

This is important for two reasons.

1. It tends to show that the translation process currently in place does not adequately benefit from past translations. Indeed, had the current translation process had a properly functioning translation memory, then it stands to reason that the same source sentence should always produce the same target sentence. However, this conclusion should be qualified: We do not know for sure at this point that the unique sentence is necessarily the original and that the multiple sentences associated with it are necessarily the translations. It could be the other way around *in some cases*, in the scenario where the original is rephrased and the translation stays the same. However, it is more than likely that our first conclusion stands, and that a TM is missing or misbehaving in the current translation pipeline.
2. The nature of the duplicated translations casts doubt on the quality of the translation process currently in place. A cursory examination of the multiple alternatives associated with the same sentence shows that some of these alternatives are simply erroneous. This is what is illustrated in the last two rows of Table 4 (`SUR EN TERRAIN MONTAGNEUX`).

R2. We recommend that a TM be added to the current translation process in order to avoid re-translating previously seen sentences. If a TM is already in place, then it should be reviewed to ensure it is operating accurately.

5 Findings on the CAP archive

During the process outlined in Section 4.1, we encountered several anomalies with the data we were sent. As explained in Section 2.2, some of these anomalies render textual data unusable in WATT (or any machine translation system for that matter).

5.1 Inconsistent newline encoding in XML files

The textual description of warnings in CAP files sometimes uses the Unix newline character (0xA) and other times the Windows newline characters (0xD 0xA). This is not a good practice, as this may create problems down the line. Figure 4 illustrates the problem. This problem was detected in more than 24,000 CAP files in the whole EC archive, but seems to only affect files produced before March 2015. Our work compensates for this inconsistency.

```
Significant snowfall amounts expected this weekend. ■
■
The first Alberta clipper like disturbance which brought some snow to all
regions [...] This is below the warning threshold of 15 cm per 12 hours. □
□
Poor winter driving conditions from low visibility in the snow and occasional
blowing snow in exposed areas are expected. ■
```

Figure 4 – Line ending incoherence from description in XML file T_WOCN12_C_CWTO_201502071051_4B2A889E.cap.
■ = Unix new line, □ = Windows new line.

R3. Harmonize line endings, by selecting Unix newline format (recommended) and making sure this choice is applied everywhere. Note that this recommendation may already be implemented.

5.2 Erroneous paragraph breaks in XML files

Much like in a regular word processor, hard returns (carriage returns) should only be used in the CAP description to denote true paragraph breaks. It is therefore troubling to find descriptions formatted like the one shown in Figure 5. This seems to be a recurrent problem affecting French descriptions primarily. The example shown is particularly problematic because every first character on the lines is capitalized, regardless of proper grammar.

We surmise that the current case-restoration process assumes that every paragraph must start with a capital letter. Since paragraph breaks are inserted where they are not needed, it follows that capital letters are added where they are not warranted.

```
Un mélange de temps hivernal affectera le Labrador à partir de la ■
Nuit de jeudi à vendredi ou de vendredi et persistera une partie de ■
La fin de semaine. ■
■
Une intense dépression traversera le Labrador en début de journée ■
Samedi, puis se déplacera pour se trouver au large de la côte jusqu'à ■
Dimanche. De la neige commencera à tomber sur l'ouest du Labrador ■
Jeudi soir et dans la nuit de jeudi à vendredi et se propagera vers ■
L'est vendredi. La neige se changera en grésil et en pluie ■
Verglaçante, puis finalement principalement en pluie d'ici tard ■
Vendredi. La pluie se changera de nouveau en neige dans la plupart ■
Des régions samedi. La neige devrait persister sur l'est du Labrador ■
Dimanche; on pourrait observer d'importantes ■
Accumulations. De plus, des vents forts se lèveront samedi et ■
Persisteront dimanche. ■
```

Figure 5 – Incorrect paragraph formatting in CAP file T_WOCN17_C_CWHX_20151104200644444.cap. The symbol ■ denotes a hard return.

We had to compensate for this by ignoring newlines and paragraph breaks altogether, and using our own sentence segmentation algorithms. Thankfully, WATT's data preparation tools ignore the case of these warnings, so the erroneously capitalized words are ignored.

However, what is concerning here is the fact that discussions like the one shown in Figure 5 were probably published as is. If this assumption is correct, most readers will have noticed this poor formatting. Even if the whitespace is reorganized, say, by being somehow reformatted on EC's website, the ungrammatical capitalization of words will have remained.

R4. Fix the problem of erroneous paragraph breaks in CAP files and adopt a consistent, principled way of delimiting paragraphs in a description. One way to do this is to use a single newline to denote the beginning of a new paragraph.

5.3 Remnants of MTCN format in the description text

The internal format of MTCN bulletins (see Figure 2) uses all uppercase words and presents a few oddities: double periods (`..`), absence of apostrophes, etc. Angle brackets are also used in this format as delimiters (e.g. `>>50>>`).

This is acceptable as an internal format, but not when those oddities appear for all Canadians to read on EC's website. Nonetheless, there appears to be remnants of those peculiarities in the description found in CAP files. The double period is present in 86 files, as illustrated in Figure 6. The angle brackets were found in 7 files.

Sentences or words needlessly written in all capitalized characters are also present in more than 1500 CAP alerts, giving rise to fragments like "RADAR indicates a west-east" or mixed-case discussions e.g. Snowfall, with total amounts of about 10 cm is expected. A LOW PRESSURE SYSTEM OVER CENTRAL MONTANA THIS AFTERNOON IS FORECAST […].

The problem may not be as important as it seems though, because the majority of these all-capped words are URLs as in "visitez le HTTP://WWW.ALBERTAHEALTHSERVICES.CA/1926.ASP.", which only rarely "breaks" the link. EC appears to be aware of the problem, since discussions include sentences like Veuillez consulter WWW.METEO.GC.CA/HURRICANE (en lettres minuscules). Nonetheless, it would make sense to have these URLs in lowercase.

A trough of low pressure will remain entrenched over the maritmes over the next several days. Rain will be prolongued and will become heavy on Monday as the trough starts to nudge toward Cape Breton Island. Some regions may have total rainfall amounts associated with the trough totalling 80 millimetres by late Monday. Of significance..Colder air in the wake of the trough will cause rain to change to snow over western parts of the island Monday afternoon. However..The ground is still relatively warm causing continuous snow melt..Thus lessening its impact.

Figure 6 – Formatting errors carried over from MTCN format in CAP description (..). In CAP file T_WOCN15_C_CWHX_2015112119351515.cap. (The words Maritimes and prolonged are also misspelled.)

R5. Investigate why there seems to be MTCN formatting artefacts (double periods, all uppercase letters, angle bracket delimiters) in the CAP descriptions in order to eliminate them.

R6. Write URLs in all lowercase letters in the CAP discussions.

5.4 Significant inconsistencies between the English and French versions

Sentence-aligning the French and English descriptions within a single CAP led us to discover that some CAP files contain a French version very different in content from the English one.

Figure 7 illustrates the problem: A single CAP file contains two dissimilar descriptions. Typically, the English version is much more complete than the French version. If we are right in thinking that these discussions were published as is, then the French speakers

consulting these problematic warnings were simply not informed of the weather conditions.

It is worth noting that these problems typically arise when the CAP file contains more than two descriptions (3 and up). We are unsure as to what gives rise to these multiple discussions, possibly rapidly changing conditions and/or a malfunctioning translation process.

French description
On ne prévoit plus de bourrasques de neige. avertissement de bourrasques de neige
English description
Snow squalls are expected. Under the snow squall bands, visibilities will be significantly reduced due to the heavy snow combined with blowing snow, and snow will quickly accumulate. Lake effect snow continues to impact portions of central Ontario today. Radar indicates that the lake effect activity is becoming more organized. Radar depicts an organization and intensification of a snow squall band near Parry Sound as well as a few intense bands near Bracebridge blowing in off Georgian Bay. These bands are expected to slowly meander southward as the day progresses. Under the snow squall bands, visibilities will be significantly reduced due to the heavy snow which will quickly accumulate. Local additional snowfall accumulations may vary from 10 to 20 centimetres by late this evening, with the most intense activity expected this afternoon and evening. Visibility will be suddenly reduced to near zero at times in heavy snow and blowing snow. Travel is expected to be hazardous due to reduced visibility. If visibility is reduced while driving, turn on your lights and maintain a safe following distance. Please continue to monitor alerts and forecasts issued by Environment Canada. To report severe weather, send an email to storm.ontario@ec.gc.ca or tweet reports to #ONStorm.

Figure 7 – The French and English descriptions within the same CAP file differ significantly. In T_WWCN11_C_CWTO_20160101211722222.cap.

There is no exhaustive way of detecting this problem in all available CAPs, short of inspecting them all manually, but a reasonable approximation is reached by finding descriptions with very dissimilar sizes within the same CAP files.

Table 6 shows the number of these alignment problems found, for each type of weather warnings. Even if the problem is not very common, it would be interesting to investigate it in order to understand what leads to these problems, especially since it keeps appearing even in recent warnings (2015). The complete list of warnings containing this problem in provided in Appendix A.

Table 6 – Alignment problems in CAP files for different bulletin types. The complete list of warnings containing this problem in provided in Appendix A.

Bulletin type	Nb of CAPs with alignment problem
WFCN	4
WOCN	18
WSCN	0
WTCN	0
WUCN	38
WWCN	25
Total	**85**

R7. Investigate why some CAP files contain a French description much shorter and less informative than the English one (see Appendix A).

5.5 Language quality

In the previous sections, we have pointed out some issues that directly influence the grammaticality and style of the weather warnings studied.

We attempt to quantify these grammar and vocabulary problems here, for each language *taken separately*. Consequently, we do not try to measure the quality of the translations produced given an original.

For each language, we created a random sample of 200 distinct sentences found in CAP descriptions, for each language. In each corpus, we identified and counted the number of various vocabulary and grammatical flaws. We used the software Antidote[7] to facilitate this task. We identify 7 types of errors, for which we counted the number of occurrences. This is shown in Table 7.

In French, we found 76 errors in total, affecting 53 distinct sentences (a single sentence can contain more than one error). This amounts to 53/200 = 26.5% of sentences. In English, this figure falls to 25 errors found in 22 different sentences, i.e. 22/200 = 11% of sentences. Examples of these flaws are provided in Table 8.

Minor typographical errors were ignored in this evaluation. For instance, the conjunction "which" should be preceded by a comma; In French, a unit like "centimeter" (cm) should be separated from the measure by a space. We did not count these (and other) errors.

[7] http://www.antidote.info/ – Antidote is a recall-oriented grammar checker whose high performance makes it an ideal tool for this task.

Table 7 – Number of different types of error for 200 randomly selected sentences in French and English, taken from the CAP descriptions. There are 4114 words in French and 3491 in English.

Error type	Nb. in French	Nb. in English
CR: case problems caused by carriage returns (Section 5.2)	40	4
G: incorrect grammar/syntax	18	1
M: MTCN format remnants (Section 5.3)	2	3
P: misspelled place name	8	3
S: very poor style	2	2
T: important typographical problem	3	4
V: misspelled word (vocabulary problem)	3	8
Total	**76 errors**	**25 errors**

Table 8 – Examples of different types of error for 200 randomly selected sentences in French and English, taken from the CAP descriptions.

Error type	Example in French	Example in English
CR	Il est trop tôt pour déterminer les répercussions exactes de cette Dépression, mais il y a actuellement un risque d'importantes Quantités de pluie et de vents forts pour l'ensemble de l'île.	From 30 to 40 millimetres of rain is Expected generally and up to 60 mm over the Eastern Townships and Beauce areas, a rainfall warning has thus been issued for these Areas.
G	Les orages se trouve du lac Hardwood	Strong gusty winds, large hail, heavy downpour and intense lightning
M	d'ici dimanche matin. ..On prévoit des vents forts	..Winter Storm likely to bring mix of snow, ice pellets and rain with strong winds Tuesday and Tuesday night.
P	Régions de Fermont, de Schefferville et de la grande iv	will track towards the maritimes on Thursday
S	Dans l'intérieur-sud, un gel rapide rendra les surfaces mouillées Glacées et glissantes sur l'intérieur-sud au cours de la nuit et Vendredi.	The fog is expected to move inland this evening and may affect the QEW for the evening commute.
T	Mardi une dépression se formera	For areas north of lakes Erie and Ontario a period of ice pellets
V	Les secteurs sous l'orage 10 km à l'ouest de Marquis ont réçu	In addition a low pressure system near the Lower Great Lakes continues to generate moderate northeasterly winds today.

The figures shown in Table 7 show that there is a quality problem in both languages. The problem is particularly acute in French. Even if we ignore the CR errors, probably due to a processing artefact, there are still 76 – 40 = 36 errors out of 4114 French words, which amounts to an error every 114 words. Since a French warning contains 102 words on average (see Table 3, page 8), on average, one could expect to see an error in 9 out of 10 bulletins in French, which is arguably very poor.

The larger number of errors in French compared to English suggests a problem in translation rather than one of creation, since the bulk of French warnings are the product of translation, in Canada. This does not mean that the creation of original warnings in English is error-free, as we show in Table 7.

The spelling of place names remains a problem, as also highlighted during the evaluation campaigns RALI participated in, for WATT, in 2014 and 2012. The case of the toponyms is almost always the problem in this case. It seems the case-restoration process could be improved. The simplest way to achieve this is to provide a list of place names to the system so that it "knows" how to spell those. RALI has implemented a more sophisticated version of this strategy in WATT.

R8. Spend effort finding a way to sanitize the description text used when creating new warnings, possibly by looking into RALI's proposition of a computer-aided tool to guide this creation. Guy Lapalme has shown a prototype for such a tool with a text-completion feature trained with existing sentences. Such a tool, when properly trained and designed, does *not* slow down the meteorologist.

R9. Use a grammar checker like Antidote or Word to check the quality of original warnings and their translation. It is a very small investment with a significant potential.

R10. Use a curated list of place names to validate the spelling and case of toponyms mentioned in warnings.

R11. Identify other ways of improving the quality of French descriptions.

6 Conclusion

This study aimed at describing the text received in the CAP archive sent by EC to RALI, with a view to using this data to improve the current version of WATT.

It turns out that:

1. There are multiple problems that negatively affect the quality of the human-produced text in these descriptions, ranging from computer problems in the processing pipeline for those warnings, to human errors when producing and translating these warnings.
2. Many of the problems identified affected the visible quality of the warnings as published on EC's website for all to consult.
3. The quality of the French text is significantly inferior to the English corpus.

It is not yet clear at this point how RALI will compensate for this poor quality when we attempt to use this data to improve WATT. We will probably have to heuristically filter out some of the material provided so as to prevent WATT from learning from unsound material.

7 Recommendations

This is a recapitulation of the recommendations we make in this document.

Appendix A – List of bulletins with inconsistent French and English discussions

We identified 85 files with suspiciously different description texts in the CAP archive (see Section 5.4). We list them here. The `.cap` file extension is omitted to simplify the list.

1. T_WFCN11_C_CWTO_20140617215 9_6B50E016
2. T_WFCN13_C_CWWG_20130716015 0_81B41AF4
3. T_WFCN13_C_CWWG_20140706013 3_63C04938
4. T_WFCN13_C_CWWG_2015070421142727
5. T_WOCN10_C_CWUL_20130913084 7_BCB51D8F
6. T_WOCN10_C_CWUL_20131027044 7_C07ADF9B
7. T_WOCN10_C_CWUL_20131127192 7_EF573337
8. T_WOCN10_C_CWUL_20140213092 6_B2AF2218
9. T_WOCN10_C_CWUL_20140319175 7_31F36D84
10. T_WOCN10_C_CWUL_20140325080 3_44F4B179
11. T_WOCN10_C_CWUL_20140423071 7_F682F2F1
12. T_WOCN10_C_CWUL_20140831075 8_1B8FB6B2
13. T_WOCN10_C_CWUL_20141008190 6_DE9D4FFF
14. T_WOCN10_C_CWUL_20141201195 2_E0E5A7B5
15. T_WOCN10_C_CWUL_20150130081 9_091B81B3
16. T_WOCN10_C_CWUL_2015032108512020
17. T_WOCN10_C_CWUL_2015050220342424
18. T_WOCN10_C_CWUL_2015050621362828
19. T_WOCN10_C_CWUL_2015071914452424
20. T_WOCN11_C_CWWG_20141223100 6_679BA182
21. T_WOCN13_C_CWWG_20141223100 6_32ED719A
22. T_WOCN14_C_CWHX_2015092408092222
23. T_WUCN11_C_CWTO_20130911222 8_AC2B6F09
24. T_WUCN11_C_CWTO_20140701172 1_65B3C1F0
25. T_WUCN11_C_CWWG_20130704003 4_7C1C293D
26. T_WUCN11_C_CWWG_20130704011 0_33F1EE2F
27. T_WUCN11_C_CWWG_20130704014 3_33AC84C7
28. T_WUCN11_C_CWWG_20130704024 4_71B3EDC9
29. T_WUCN11_C_CWWG_20130704032 4_2DBDF59C
30. T_WUCN11_C_CWWG_20130704034 9_FEC647BE
31. T_WUCN11_C_CWWG_20130704061 4_7D81E277
32. T_WUCN11_C_CWWG_20130704080 1_4DBF57D6
33. T_WUCN11_C_CWWG_2015061302241515
34. T_WUCN11_C_CWWG_2015071300241717
35. T_WUCN12_C_CWWG_20130704003 4_A2A70B71
36. T_WUCN12_C_CWWG_20130704011 0_499FD8C2
37. T_WUCN12_C_CWWG_20130704013 4_9F68011A
38. T_WUCN12_C_CWWG_20130704024 4_2AE28895
39. T_WUCN12_C_CWWG_20130712031 2_3C9C1ABE
40. T_WUCN13_C_CWVR_20130905015 3_EBC13B28
41. T_WUCN13_C_CWWG_20130430000 6_5B8CA174
42. T_WUCN13_C_CWWG_20130527043 4_F2FE71CF
43. T_WUCN13_C_CWWG_20130630225 5_FEDBAC26
44. T_WUCN13_C_CWWG_20130704014 3_A5FAB1B8
45. T_WUCN13_C_CWWG_20130704034 3_8A8DCA27
46. T_WUCN13_C_CWWG_20130712024 0_921A8E94
47. T_WUCN13_C_CWWG_20130814225 1_CAB6F7E1
48. T_WUCN13_C_CWWG_20130831005 7_FAE3628F
49. T_WUCN13_C_CWWG_2015061300342828
50. T_WUCN14_C_CWHX_20130911215 3_8242C45C
51. T_WUCN15_C_CWWG_20130629232 5_94CB6FAB
52. T_WUCN15_C_CWWG_20130630010 1_C11DEAC3
53. T_WUCN15_C_CWWG_20130630025 1_E94EF764
54. T_WUCN15_C_CWWG_20130630033 3_C7CBAE88
55. T_WUCN15_C_CWWG_20130630035 0_5D9EE26C
56. T_WUCN15_C_CWWG_20130630044 4_103FA618
57. T_WUCN15_C_CWWG_20130704200 9_3B7AAFCE
58. T_WUCN15_C_CWWG_20130712030 8_352B878B
59. T_WUCN15_C_CWWG_20140710044 5_1688D1CE
60. T_WUCN16_C_CWWG_20130704200 9_32E41386
61. T_WWCN10_C_CWUL_20130515084 1_DD524432
62. T_WWCN10_C_CWUL_20130706203 6_D7D9E6D4
63. T_WWCN11_C_CWHX_20150216084 9_AE538FBA
64. T_WWCN11_C_CWTO_20130522192 4_548FF9B1
65. T_WWCN11_C_CWTO_2016010109401919
66. T_WWCN11_C_CWTO_2016010121172222
67. T_WWCN11_C_CWVR_20131101114 4_3696EAA7
68. T_WWCN11_C_CWVR_20140111123 1_C94F40A6
69. T_WWCN11_C_CWWG_20130125034 6_3955207E
70. T_WWCN11_C_CWWG_2015081300083333
71. T_WWCN11_C_CWWG_2015082222531010
72. T_WWCN12_C_CWTO_20130531223 9_93EB26E6
73. T_WWCN12_C_CWTO_20150116032 0_3E5260DA
74. T_WWCN13_C_CWNT_20151225204020 20
75. T_WWCN13_C_CWVR_20131217142 1_623BC577
76. T_WWCN13_C_CWVR_20131227054 9_8BBAAA2A
77. T_WWCN13_C_CWVR_2015122709232828
78. T_WWCN13_C_CWWG_20140705235 0_90D03B38
79. T_WWCN14_C_CWWG_2015122716173030
80. T_WWCN16_C_CWHX_20141103084 2_7577CEE3
81. T_WWCN16_C_CWNT_20131226092 3_629F34CD
82. T_WWCN16_C_CWWG_20130319214 4_8BD3FBDC
83. T_WWCN16_C_CWWG_20130414035 5_6C3B59EE
84. T_WWCN18_C_CWVR_2015122709232828
85. T_WWCN19_C_CWVR_20130419223 9_4D8A98EF