A Corpus-based Approach to Weather Report Translation

Thomas Leplus, Philippe Langlais and Guy Lapalme
RALI-DIRO, Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec,
Canada H3C 3J7
{leplusth,langlais,lapalme}@iro.umontreal.ca

Abstract

We describe a corpus-based rational reconstruction of a landmark application of machine translation, the Canadian English to French weather report translation system. This system, which has been in operation for more than 20 years, has been developed using a classical symbolic approach. We describe our experiment in developing an alternative approach based on the analysis of hundreds of thousands of weather reports. We show that it is possible to obtain excellent translations using statistical and translation memory techniques.

1 Introduction

In the mid seventies, a group of linguists and computer scientists of Université de Montréal (*TAUM* group) developed an English to French machine translation system for weather reports which became known as TAUM-MÉTÉO described in (Hutchins and Somers, 1992, chapter 12). It involves three major steps: a dictionary look-up, a syntactic analysis and a light syntactic and morphological generation step.

The transfer from English to French is encoded at the word level into three special purpose lexicons: idioms (e.g. *blowing snow* ↔ *poudrerie*), locations (e.g. *Newfoundland* ↔ *Terre Neuve*) and a general dictionary containing syntactic and semantic features (e.g. *amount*=N((F,MSR),*quantite*) which means that *amount* translates into the feminine F French noun N *quantite* which is a measure MSR noun).

The syntactic stage is the result of a detailed analysis that was done by hand at the early stage of the prototype. (Chandioux, 1988) reports that MÉTÉO-2, a subsequent system that became operational at Environment Canada, used fifteen different grammars categorized into five major types from which the syntactic analysis chooses the most appropriate one.

The third and last step performs French word reordering (e.g. adjectives are placed after the noun they modify), preposition selection (e.g. we say ***à*** *Montréal* but ***en*** *Nouvelle-Écosse* and ***au*** *Manitoba*) plus a few morphological adjustments (e.g. *le été* → *l'été*).

This system has been in continuous use since 1984 translating up to 45,000 words a day. (Grimaila and Chandioux, 1992) argues that one of the reasons for the success of the MÉTÉO system is the nature of the problem itself: a specific domain, with very repetitive texts and particularly unappealing to translate by a human (see for example the reports shown in Figure 1). Furthermore, the life of a weather report is, by nature, very short (approximatively 6 hours), which make them an ideal candidate for automation.

Professional translators can be prompted to correct machine output when the input English text cannot be parsed, often because of spelling errors in the original English text. It is one of the very few Machine Translation system in the world from which the unedited output is used by the public in everyday life without any human revision.

Some alternatives to Machine Translation (MT) have been proposed for weather reports, namely multilingual text generation directly from raw weather data: temperatures, winds, pressures etc. These generation systems also need some human template selection for organizing the report. Generating text in many languages from one source is quite appealing from a conceptual point of view and has been cited as one of the potential applications for natural language generation (Reiter and Dale, 2000); some systems have been developed (Kittredge et al., 1986; Goldberg et al., 1994; Coch, 1998) and tested in operational contexts. But until now, none has been used in every day production at the same level as the one attained by MT. One of the reasons is that meteorologists prefer to

write their reports in natural language rather than selecting text structure templates.

Given that recent statistical and corpus approaches for machine translation have proven their value in some contexts, we decided to see how well these approaches would fit in the context of the weather report translation. We obtained from Environment Canada 309,531 forecast reports in both French and English produced during 2002 and 2003. The current reports are available on the web at `http://meteo.ec.gc.ca/forecast/textforecast_f.html`. This site is continually being updated. We used this corpus as a source for developing a new system for weather reports and thus gave *rebirth* to one of the most successful machine translation system until now.

We describe in section 2 the data we received and what preprocessing we performed to obtain our MÉTÉO bitext. We present in section 3 the first prototype we devised and report on its evaluation in section 4. We finally discuss this work in section 5.

2 The corpus

As our system is based on both memory and statistical based approaches, the first thing we need is a bitext i.e. an aligned corpus of corresponding sentences in French and English weather reports. Like all work on real data, this conceptually simple task proved to be more complicated than we had initially envisioned. This section describes the major steps of this stage.

2.1 The raw corpus

We received from Environment Canada files containing both French and English weather forecasts produced during 2002 and 2003. Both the source report, usually in English, and its translation, produced either by a human or by the current MÉTÉO system, appear in the same file. One file contains all reports issued for a single day. A report is a fairly short text, on average 304 words, in a telegraphic style: all letters are capitalized and non accented and almost always without any punctuation except for a terminating period. As can be seen in the example in Figure 1, there are few determiners such as articles (*a* or *the* in English, *le* or *un* in French). A report usually starts with a code identifying the source which issued the report. For example, in `FPCN18 CWUL 312130`, `312130` indicates that the report was produced at 21h30 on the 31st day of the month; `CWUL` is a code corresponding to Montreal and the western area of Quebec. A report (almost always) ends with a closing markup: `END` or `FIN` according on the language of the report; if the author or the translator is a human, his or her initials are added after a slash following the markup.

2.2 Matching English and French reports

We first determine the beginning and end of each weather forecast using regular expressions to match the first line of a forecast, which identifies the source which issued it, and the last line which usually starts with `END` or `FIN`.

Then we distinguish the English forecasts from the French ones according to whether they ended with END or FIN. Given the fact that we started with a fairly large amount of data, we decided to discard any forecast that we could not identify with this process. We were left with 270,402 reports.

We also had to match English and French forecasts that are translations of each other. As we see in Figure 1, the first line of the two reports is almost the same except for the first part of the source identifier which is `FPCN18` in English and `FPCN78` in French. After studying the data, we determined that this shift of 60 between English and French forecasts identifiers seemed valid for identifiers from `FPCN10` through `FPCN29`. These identifiers being the most frequent, we decided to keep only these in our final bitext.

This preprocessing stage required about 3,000 lines of Perl and Bash scripts and a few weeks of monitoring. Out of the 561 megabytes of text we received, we were left with *only* 410 megabytes of text, representing 127,950 weather report pairs.

2.3 Creating a bitext

To get a bitext out of this selected material, we first segmented automatically the reports into words and sentences using an in house tool that we did not try to adapt to the specificity of the weather forecasts. Actually, this segmenter did not always manage to identify end of sentences accurately.

We then ran the Japa sentence aligner (Langlais et al., 1998) that took around 2 hours of a desktop workstation time to identify 4.1 million pairs of sentences from which we removed about 44 000 (roughly 1%) which were not one to one sentence pairs.

FPCN18 CWUL 312130

SUMMARY FORECAST FOR WESTERN QUEBEC ISSUED BY ENVIRONMENT CANADA

MONTREAL AT 4.30 PM EST MONDAY 31 DECEMBER 2001 FOR TUESDAY 01 JANUARY 2002. VARIABLE CLOUDINESS WITH FLURRIES. HIGH NEAR MINUS 7.

END/LT

FPCN78 CWUL 312130

RESUME DES PREVISIONS POUR L'OUEST DU QUEBEC EMISES PAR ENVIRONNEMENT CANADA

MONTREAL 16H30 HNE LE LUNDI 31 DECEMBRE 2001 POUR MARDI LE 01 JANVIER 2002. CIEL VARIABLE AVEC AVERSES DE NEIGE. MAX PRES DE MOINS 7.

FIN/TR

Figure 1: An example of an English weather report and its French translation.

We divided this bitext into three non overlapping sections as reported in Table 1: TRAIN (January 2002 to October 2003) for training the translation and language models, BLANC (December 2003) for tuning them and TEST (November 2003) for testing.

The TEST section was chosen to be different from the TRAIN period in order to recreate as much as possible the working environment of a system faced with the translation of new weather forecasts.

		English		French	
corpus	pairs	words	tok.	words	tok.
TRAIN	3,933	30,173	9.2	36,895	10.9
BLANC	115	884	2.8	1,082	3.3
TEST	34	268	1.8	330	2.0
total	4,081	31,325	9.4	38,307	11.2

Table 1: Main characteristics of the subcorpora used in this study in terms of number of pairs of sentences, English and French words and tokens (different words). Figures are reported in thousands.

2.4 Characteristics of the bitext

A fast inspection of the bitext reveals that sentences are fairly short: an average of 7.6 English and 8.4 French words. Figure 2 shows the length distribution for the three sections of our bitext. Most sentences are repeated, only 8% of the English sentences being hapax legomena. About 90% of the sentences to be translated can be retrieved with at most one edit operation i.e. insertion, deletion or substitution of a word. These properties of our bitext naturally call for a memory based approach for translating weather reports.

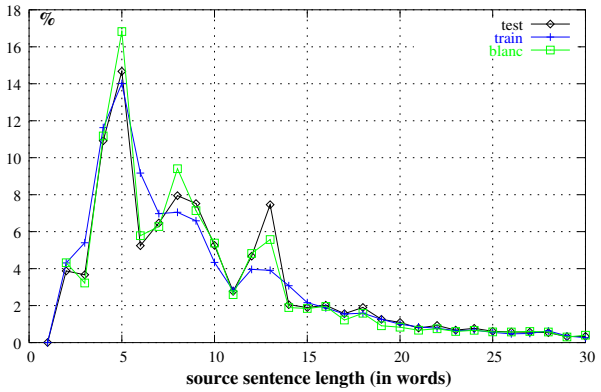

Figure 2: Distribution length of the identified English sentences.

3 The translation system

The strategy we devised to build this translation system is a combination of memory and statistic-based approaches. The scenario for translating a sentence is the following: the source sentence is preprocessed in order to handle more easily entities such as dates, hours or numbers (see section 3.1). Sentences which are closest to this processed source sentence are retrieved from the translation memory. If the sentences found involve few or no edit operations (a notion we quantify in section 3.2), we return the most frequent translation found (postprocessed to remove the meta-tokens introduced by the preprocessing stage); in cases where many edit operations are required, the source sentence is fed into a statistical machine translation (SMT) engine. Figure 3 illustrates this process for the translation of one sentence.

3.1 Preprocessing

Before the creation of the memory, eight classes of tokens (*punctuation*, *telephone numbers*, *months*, *days*, *time*, *numeric values*, *range of values* and *cardinal coordinates*) are identified with regular expressions at the character level

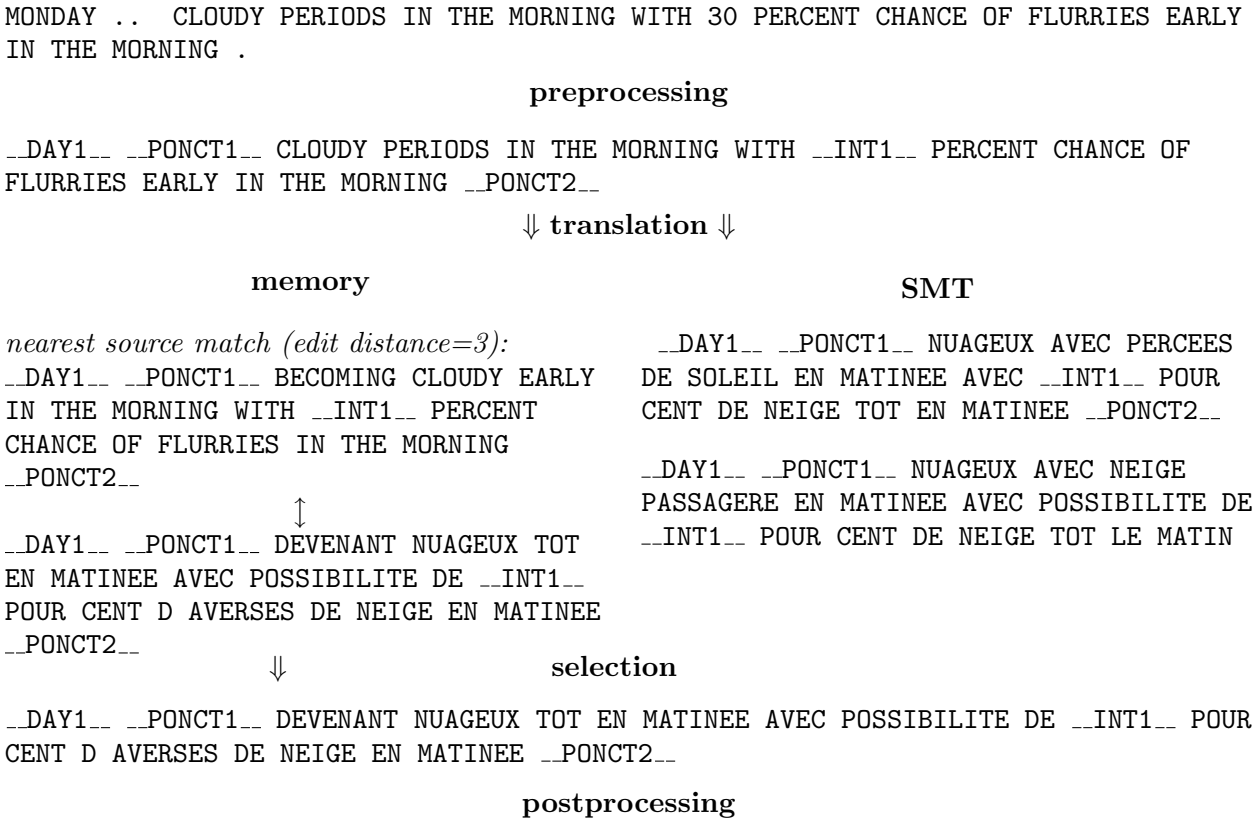

Figure 3: Illustration of a translation session. The source sentence is first preprocessed. Both the memory and the SMT engine are queried. Here, the memory found only an approximative match with an edit distance of 3. The two first translations produced by the SMT engine are reported in the right column. A translation is then selected and postprocessed.

and replaced by a corresponding tag (numbered if there are more than one of the same class in a sentence) in the bitext. This process is different from the creation of specialized lexicons used in the current MÉTÉO system. We only used automatic tools for building the system.

3.2 The translation memory

First, the size of the translation memory affects our system in two ways. If we store only the few most frequent English sentences and their French translations, the time for the system to look for entries in the memory will be short. But, on the other hand, it is clear that the bigger the memory, the better will our chances be to find sentences we want to translate (or ones within a short edit distance) even if these sentences were not frequent in the training corpus. We can see in Figure 4 that the percentage of sentences to translate found directly in the memory grows logarithmically with the size of the memory until it reaches approximately 20,000 sentences. With a memory of 489,000 sentences, we can obtain a peek of 83% of sentences found in the memory.

The second parameter for the translation memory is the number of French translations stored for each English sentence. Among the 488,792 different English sentences found in our training corpus, as many as 437,418 (89%) always have the same translation. This is probably because most of the data we received from Environment Canada is actually machine trans-

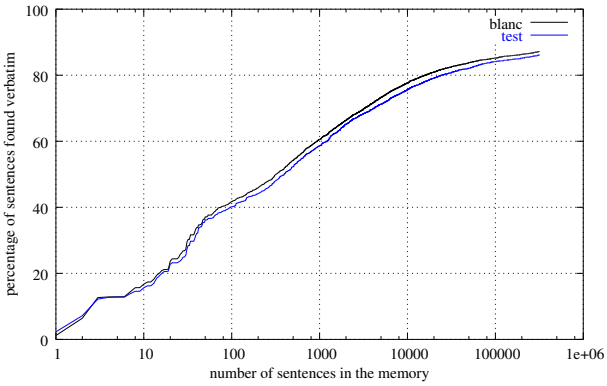

Figure 4: Percentage of the TEST and BLANC sentences found verbatim as a function of the number of sentences kept in the memory.

lated and has not been edited by human translators. So we decided to store in the memory at most five translations per sentence (i.e. for the 11% of sentences with several translations, we kept only the five most frequent ones).

3.3 Postprocessing

The output of the translation memory is postprocessed in order to transform the meta-tokens introduced during preprocessing into their appropriate word form. We observed on a held-out test corpus that this cascade of pre and postprocessing (illustrated in Figure 3) increased the coverage of the memory while significantly reducing its size: with a memory of 296,000 sentences with meta-tokens (40% less than without meta-tokens), we found 87% of the sentences to translate in the memory.

3.4 The translation engine

We used an in house implementation of the decoder described in (Nießen et al., 1998). This decoder relies on an IBM translation model 2 (Brown et al., 1993) as well as on an interpolated ngram language model. We used both the TRAIN and BLANC section of our bitext (without preprocessing) to train both models.

We must share with the reader the singular and privileged moment we had monitoring the training. A few minutes on an ordinary desktop computer were enough to make us feel what the life of computer linguist will be in the (near?) future: the training translation perplexity was as low as 4.7 and the language model ones were of 4.6 for a 3-gram model. Just to give one example, when we train a trigram language model on the Canadian Hansards, we observe a perplexity of 65, which is already very low compared to perplexities measured on more general corpora. This extremely low perplexity can be explained by the small vocabulary of the MÉTÉO corpus, as well as the clear dominance of pattern of sentences (e.g `THE NEXT FORECASTS WILL BE ISSUED AT...`), which is accentuated by the fact that the translation produced are mostly non post-edited machine output.

4 Results

We report in Table 2 the performance of the SMT engine measured on the TEST corpus in terms of Word Error Rate (normalized Levenshtein distance) and Sentence Error Rate (percentage of the system's translations that does not match exactly the reference). These measurements are provided for two scenarios: one in which the system produces a single translation (1-best) and one in which the system produces at most five translations. In the latter case, the score for a source sentence is computed using an oracle which picks the best translation among the ones proposed[1].

sent. length	sent. count	1-best WER	1-best SER	5-best WER	5-best SER
0- 5	12981	7.5	24.7	5.0	15.2
0-10	26271	18.8	52.5	14.8	42.8
0-15	31706	20.0	59.3	16.0	50.4
0-20	33932	20.8	61.7	17.0	53.3
0-25	35054	21.5	62.9	18.0	54.8
0-30	35740	22.0	63.7	18.6	55.6
all	36228	22.9	64.1	19.6	56.2

Table 2: Performance of the statistical engine on the TEST corpus (36,228 sentences) as a function of the sentence length.

Given the nature of the task, the error rates reported in Table 2 are relatively low compared to other translation tasks. On Hansard translation sessions, we usually observe with the same engine a WER around 60%, and a SER in the range of 80-90%. We found that many translations were well-formed, but with sometimes a slightly different meaning than the original source sentence. For example, in Figure 3, the translations produced by the SMT engine, *cloudy periods in the morning* is first translated as *nuaugeux avec percee de soleil en matinee*

[1]Minimization of WER was used to simulate the oracle in this experiment.

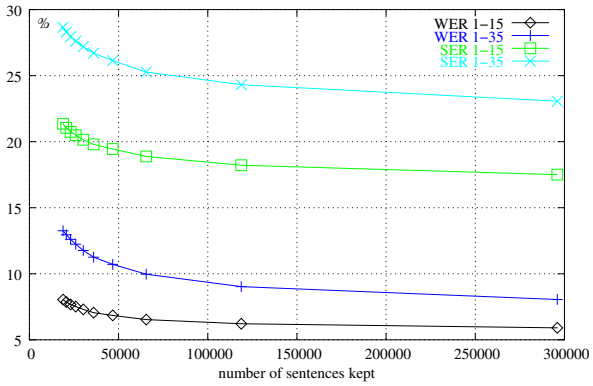

Figure 5: Performance of the memory-based approach as a function of the number of sentences kept in the memory. Each point corresponds to a frequency threshold (from 10 to 1) we considered for filtering in the training sentences. A threshold of 1 (the last dot) means that all the TRAIN bitext was considered as a memory.

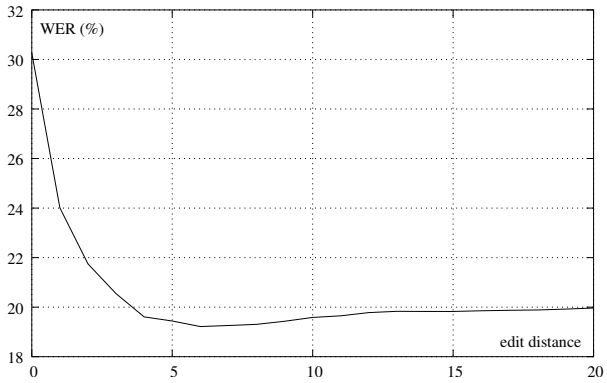

Figure 6: Performance of the prototype as a function of the threshold over which the SMT engine is queried. Only those sentences that are not verbatim in the TRAIN bitext are considered.

system	WER	SER	NIST	BLEU
memory	8.42	23.43	10.96	0.877
engine	22.88	64.14	9.07	0.620
prototype	8.22	23.42	11.00	0.877

Table 3: Performance of the memory-based approach, the statistical engine and both combined into our prototype. We computed this scores on a single reference with the mteval program (version v11a) available at `http://www.nist.gov/speech/tests/mt/`.

(cloudy with sunny periods in the morning) and the next time as *nuageux avec neige passagere en matinee* (cloudy with occasional snow in the morning). A possible explanation lies in our bitext which contains several very frequent sentences that dominate the probability mass given to the ngrams (a huge unbalanced corpus is not necessarily a good one). We still observe the usual increases in the error rates with the length of the sentences to be translated.

We can observe in Figure 5 the performance of the memory-based approach as a function of the number of sentence kept in the memory. Clearly the larger the memory is, the better the performance. The error rates reported are much lower than the SMT ones, even when the memory contains only the 5,000 most frequent sentences of the TRAIN corpus.

Figure 6 shows the translation performance we obtain by mixing the largest memory and the SMT. Since 87% of the sentences to be translated can be found verbatim in this memory, we considered only the 13% remaining ones. We observe the best performance when the SMT is only queried for sentences with an edit distance bigger than 6 to any sentence in the memory. In this configuration our prototype's WER is 19%. Finally when we keep the best configuration (threshold of 6) but translate all the sentences of our TEST corpus, our prototype performance is of 8% WER and 23% SER (Table 3). In comparison the MÉTÉO-1 system was credited with a SER of 20% (Chandioux, 1988).

5 Discussion

In this paper, we described the *re*creation of the MÉTÉO system nearly 30 years after the birth of the first prototype at the same university. The main difference between these two systems is the way they were developed. The original system is a carefully handcrafted one based on a detailed linguistic analysis, whilst ours simply exploits a memory of previous translations of weather forecasts that were, of course, not available at the time the original system was designed. Computational resources needed for implementing this corpus based approach are also much bigger than what was even imaginable when the first MÉTÉO system was developed. We show that a particularly simple system mixing a memory-based approach and a statistical one can, without almost any human effort, reproduce a quality comparable to the one produced by the current system.

This prototype can be improved on many aspects. First, we could consider a more elaborate translation model than the one we used here such as an IBM model 4 ((Brown et al., 1993)).

Another interesting candidate for which we are currently devising a decoder is a phrase based model (Koehn et al., 2003). We would like to see whether the training of such a model would replace partially or completely the memory we built in this experiment, especially since the average sentence length in our bitext is about 7 words. Our translation memory implementation is by far crude compared to the current practice in example-based machine translation (Carl and Way, 2003), since it only returns verbatim the most frequent translation found. While this make sense when the source sentence is found verbatim in the memory, it sounds a little bit odd to do so in the other cases. We are currently investigating the use of word alignment to better exploit our memory.

Our bitext thus puts us into a very particular and pleasing situation where we have enough highly specific data which is reflected by an efficient usage of a memory of previous translations. We have shown that a word-based statistical model can capture part of the information available in the memory. On a more polemic stance, the availability of such a huge quantity of previous translations gives us an opportunity to explore whether current SMT is much more than memorizing previous translations.

6 Acknowledgements

We thank Rick Jones and Marc Besner from the Dorval office of Environment Canada who provided us with the corpus of bilingual weather reports. We are also indebted to Elliot Macklovitch who provided us with articles describing the MÉTÉO system that we could not have found otherwise and to George Foster for his kind assistance in using some of his programs. This work has been supported by grants from NSERC and FQRNT.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Michael Carl and Andy Way, editors. 2003. *Recent Advances in Example-based Machine Translation*. Kluwer Academic Publishers, Doordrecht, Netherlands.

John Chandioux. 1988. Meteo (tm), an operational translation system. In *RIAO*.

J. Coch. 1998. Interactive generation and knowledge administration in multimeteo. In *Ninth International Workshop on Natural Language Generation*, pages 300–303, Niagara-on-the-lake, Ontario, Canada.

G. Foster, S. Gandrabur, P. Langlais, E. Macklovitch, P. Plamondon, G. Russell, and M. Simard. 2003. Statistical machine translation: Rapid development with limited resources. In *Machine Translation Summit IX*, New Orleans, USA, sep.

E. Goldberg, N. Driedger, and R. Kittredge. 1994. Using natural language processing to produce weather forecasts. *IEEE Expert 9*, 2:45–53, apr.

Annette Grimaila and John Chandioux, 1992. *Made to measure solutions*, chapter 3, pages 33–45. J. Newton, ed., Computers in Translation: A Practical Appraisal, Routledge, London.

W. John Hutchins and Harold L. Somers. 1992. *An Introduction to Machine Translation*. University Press Cambridge.

R. Kittredge, A. Polguère, and E. Goldberg. 1986. Synthesizing weather reports from formatted data. In *11th. International Conference on Computational Linguistics*, pages 563–565, Bonn, Germany.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Second Conference on Human Language Technology Reasearch (HLT)*, pages 127–133, Edmonton, Alberta, Canada, May.

Philippe Langlais, Michel Simard, and Jean Véronis. 1998. Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, Montréal, Canada, August.

S. Nießen, S. Vogel, H. Ney, and C. Tillmann. 1998. A dp based search algorithm for statistical machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING) 1998*, pages 960–966, Montréal, Canada, August.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press. 270 p.