

# Prodigy-METEO: Pre-Alpha Release Notes (Nov 2009)

Anja Belz

November 17, 2009

## Contents

<b>1</b>	<b>Overview of Prodigy-METEO contents</b>	<b>2</b>
<b>2</b>	<b>Inputs and human-authored forecast texts (outputs)</b>	<b>2</b>
2.1	Corpus forecast texts (outputs) . . . . .	2
2.2	Wind data (inputs) . . . . .	3
2.3	Additional human-authored forecast texts . . . . .	3
<b>3</b>	<b>System-generated forecast texts (outputs)</b>	<b>3</b>
3.1	SumTime-Hybrid . . . . .	3
3.2	Trainable systems . . . . .	5
3.2.1	PCFG Systems . . . . .	5
3.2.2	PSCFG Systems . . . . .	5
3.2.3	PBSMT Systems . . . . .	6
<b>4</b>	<b>Recreating the complete 5-fold version of Prodigy-METEO</b>	<b>6</b>
<b>5</b>	<b>Example Use of Prodigy-METEO</b>	<b>7</b>
<b>6</b>	<b>Contacts</b>	<b>7</b>

# 1 Overview of Prodigy-METEO contents

This is the documentation for the pre-alpha release of the Prodigy-METEO Corpus. The contents are structured as follows:

Prodigy-METEO_pre-alpha	top-level directory
data/	
inputs/	wind data
outputs/	wind forecast texts
human-authored/	
corpus/	wind forecast texts extracted from the original SumTime-METEO forecast files
from-data/	wind forecast texts produced by 3 new meteorologists in the standard way directly from original SumTime-METEO wind data
rewritten/	wind forecast texts produced by 3 new meteorologists by rewriting (improving) original SumTime-METEO forecast files
system-generated/	
SumTime-Hybrid/	wind forecast texts generated by hand-crafted SumTime-Hybrid system (Reiter et al.)
trainable-systems/	wind forecast texts generated by trainable systems: 5 PCFG systems, 2 PSCFG systems, and 2 PBSMT systems; each system exists in 5 versions, each trained on the training set in one of the 5 folds
PBSMT/	
fold1/	
fold2/	
fold3/	
fold4/	
fold5/	
PCFG/	
fold1/	
fold2/	
fold3/	
fold4/	
fold5/	
PSCFG/	
fold1/	
fold2/	
fold3/	
fold4/	
fold5/	
doc/	
Prodigy-METEO.pdf	detailed documentation on all aspects of Prodigy-METEO (this document)
belz-kow-enlg09.pdf	paper describing the trainable systems included in this release; Belz & Kow, 2009, System Building Cost vs. Output Quality in Data-to-Text Generation, ENLG'09.
scripts/	
normalise-for-human-evaluation.pl	script for normalising wind forecast text before evaluation by human evaluators
normalise-for-metric-evaluation.pl	script for normalising wind forecast text before evaluation by automatic metrics

## 2 Inputs and human-authored forecast texts (outputs)

Prodigy-METEO-beta/data/

The main components of the Prodigy-METEO corpus are a set of pairs of wind data (input) and corresponding wind forecast text (output) which can be used for training/building generation systems; outputs from 10 systems that were built with this corpus; and some additional human-authored wind forecasts.

The idea is to release the data and other information necessary to reproduce the results for the METEO data we have reported [2, 3], and to create new METEO systems and directly compare them to existing systems.

### 2.1 Corpus forecast texts (outputs)

Prodigy-METEO-beta/data/outputs/human-authored/Corpus/\*.prn.1

The Prodigy-METEO data was extracted from the SumTime-METEO corpus [6]. We used only those forecasts from SumTime-METEO that are for the period 06:00–24:00 GMT. These are issued in the a.m. and make up roughly half of SumTime-METEO). An example forecast file (5Oct2000\_03.prn) is shown in Figure 1. We then extracted all ‘wind statements’ except the one for the long range outlook (i.e. all wind forecasts statements in Figure 1 under points 2, 3 and 4, for 10m and 50m. This produced the following files for the example in Figure 1:

```

5Oct2000_03.prn.1 SSW 16-20 GRADUALLY BACKING SSE THEN FALLING VARIABLE 04-08 BY LATE EVENING
5Oct2000_03.prn.2 SSW 20-26 GRADUALLY BACKING SSE THEN FALLING VARIABLE 08-12 BY LATE EVENING
5Oct2000_03.prn.3 VARIABLE 04-08 SOON NNW INCREASING 12-16 BY MIDDAY THEN GRADUALLY BACKING W'LY
06-10 BY LATE EVENING
5Oct2000_03.prn.4 VARIABLE 08-12 SOON NNW INCREASING 15-20 BY MIDDAY THEN GRADUALLY BACKING W'LY
10-14 BY LATE EVENING
5Oct2000_03.prn.5 W'LY 06-10 SOON BACKING SSW INCREASING 34-38 BY LATE AFTERNOON EASING 30-34 BY
LATE EVENING
5Oct2000_03.prn.6 W'LY 10-14 SOON BACKING SSW INCREASING 43-48 BY LATE AFTERNOON EASING 38-43 BY
LATE EVENING

```

In this release, as in our recent work [2, 3], only the first wind statement from each a.m. forecast (with the extension .1) has been included.

## 2.2 Wind data (inputs)

```
data/inputs/*.num.1
```

The wind data inputs are vectors of time stamps and wind parameters, and were ‘reverse-engineered’, by automatically aligning wind speeds and wind directions in the forecasts with time-stamps in the wind data file (the wind data file for 5 Oct 2000 is shown in Figure 2). In order to do this, wind speed and directions in the data file have to be matched with those in the forecast. This was not straightforward, because often there is no exact match in the data file for the wind speeds and directions in the forecast. The strategy adopted was the same as in the SUMTIME work, in order to make the systems comparable.<sup>1</sup>

From each alignment, numerical data vectors were automatically created; e.g. the following is the input vector for output 5Oct2000\_03.prn.1 (the input filename is 5Oct2000\_03.num.1):

```
[[1, _SSW, 16, 20, -, -, 0600], [2, _SSE, -, -, -, -, -1], [3, _VAR, 04, 08, -, -, 0000]]
```

The input vector is a sequence of 7-tuples  $\langle i, d, s_{min}, s_{max}, g_{min}, g_{max}, t \rangle$  where  $i$  is the tuple’s ID,  $d$  is the wind direction,  $s_{min}$  and  $s_{max}$  are the minimum and maximum wind speeds,  $g_{min}$  and  $g_{max}$  are the minimum and maximum gust speeds, and  $t$  is a time stamp (indicating for what time of the day the data is valid).

In order to obtain the input vectors, we chunked the forecast texts into adjacent phrases, each of which realises one 7-tuple. Each forecast corresponds to at least one 7-tuple. One or more parts of a 7-tuple may not be realised in a given forecast. A  $-1$  value for a timestamp  $t$  means that the procedure described above failed to identify a time for a segment.

A ‘-’ value means that the corresponding wind information is not included in the forecast text.

## 2.3 Additional human-authored forecast texts (outputs)

```
data/outputs/human-authored/from-data/
data/outputs/human-authored/rewritten/
```

In collaboration with Ehud Reiter (e.g. [5]) we created additional sets of human-authored forecast texts. We asked three meteorologists (labelled humanA, humanB and humanC in the file extensions) who had not contributed to the SumTime-METEO corpus to write forecasts for 20 dates by rewriting (improving) the corpus forecasts; we asked the same three meteorologists to write new forecasts from scratch (looking just at data files) for a set of 18 different dates.

## 3 System-generated forecast texts (outputs)

```
data/outputs/system-generated/
```

### 3.1 SumTime-Hybrid

```
data/outputs/system-generated/SumTime-Hybrid/*.hyb.1
```

The rule-based SumTime system (Reiter et al.) has two modules: a content-determination module and a microplanning and realisation module. It can be run without the content-determination module, taking content representations (7-tuple sequences as described above) as inputs, and is then called SumTime-Hybrid. SumTime-Hybrid is a traditional deterministic rule-based generation system.

<sup>1</sup>Reiter *et al.* selected the time stamp of the data in the table that most closely matched the data in the forecast, and if there was not a close enough match, they derived a time stamp from the time expression in the forecast, and finally, if that could not be done with enough confidence, then time was left unspecified.

OCEANROUTES SPECTRAL WAVE AND WEATHER FORECAST.  
DUTY FORECASTER AVAILABLE AT ALL TIMES.PHONE ABERDEEN [[[phone number]]]  
FORECAST FOR:-

[[[oil1]]]/[[[oil2]]]/[[[oil3]]] FIELDS

1.INFERENCE 0300 GMT, THURSDAY, 05-Oct 2000  
LOW (989 MB),JUST SW OF THE FAEROES,WILL DRIFT SLOWLY SE TO LIE  
OVER THE EAST SHETLAND BASIN (1008 MB) BY LATE THIS EVENING THEN  
DRIFTING NE ALONG THE NORWEGIAN COAST FRIDAY MORNIMNG,FILLING.  
A TRANSIENT RIDGE OF HIGH PRESSURE WILL CROSS THE NORTH SEA FRIDAY  
AFTERNOON/EVENING AHEAD OF AN OCCLUDING FRONTAL SYTSEM WHICH WILL  
AFFECT THE NORTH SSE LATE SATURDAY MORNING ONWARDS.

[[[forecaster1]]]

2.FORECAST 06-24 GMT, THURSDAY, 05-Oct 2000

====WARNINGS: RISK THUNDERSTORM. =====

WIND(KTS) CONFIDENCE: HIGH  
10M: SSW 16-20 GRADUALLY BACKING SSE THEN FALLING  
VARIABLE 04-08 BY LATE EVENING  
50M: SSW 20-26 GRADUALLY BACKING SSE THEN FALLING  
VARIABLE 08-12 BY LATE EVENING  
WAVES(M) CONFIDENCE: HIGH  
SIG HT: AROUND 3.0 FALLING 2.0-2.5 BY MID AFTERNOON  
LATER MAINLY AROUND 2.5  
MAX HT: AROUND 5.0 FALLING 3.0-4.0 BY MID AFTERNOON  
LATER MAINLY AROUND 4.0  
PER(SEC): SEAS: 05-06. SWELL: 09 LATER 08  
WEATHER: CLOUDY WITH PATCHY RAIN AT FIRST,SOON BREAKING  
PARTLY CLOUDY/CLOUDY WITH SCATTERED SHOWERS,  
RISK THUNDERSTORM.  
VIS(NM): 3-5 IN FRONTAL MIST/RAIN,SOON IMPROVING 10+  
AROUND MIDDAY BUT FALLING 3-5 IN SHOWERS  
TEMP(C): 10-12 LATER 08-12  
CLOUD: 4-6 ST 500-700 6-8 SC 1200-1800 SOON BREAKING  
(OKTAS/FT) 4-6 CUSC 1500-2500 LOWERING 500-1000 IN SHOWERS  
WITH OCCASIONAL CB 800

LIGHTNING RISK: NIL RISING HIGH (60-80 PER CENT) BY LATE AFTERNOON

3.FORECAST 00-24 GMT, FRIDAY, 06-Oct 2000  
WIND(10M): VARIABLE 04-08 SOON NNW INCREASING 12-16 BY MIDDAY  
THEN GRADUALLY BACKING W'LY 06-10 BY LATE EVENING  
(50M): VARIABLE 08-12 SOON NNW INCREASING 15-20 BY MIDDAY  
THEN GRADUALLY BACKING W'LY 10-14 BY LATE EVENING  
SIG WAVE: AROUND 2.5 FALLING SLOWLY 1.0-1.5 BY LATE EVENING  
MAX WAVE: AROUND 4.0 FALLING SLOWLY 1.5-2.5 BY LATE EVENING  
WEATHER: SCATTERED SHOWERS,RISK THUNDER AT FIRST,  
DYING OUT BY LATE EVENING  
VIS: GOOD EXCEPT IN ANY THUNDERY SHOWERS

4.FORECAST 00-24 GMT, SATURDAY, 07-Oct 2000  
WIND(10M): W'LY 06-10 SOON BACKING SSW INCREASING 34-38 BY  
LATE AFTERNOON,EASING 30-34 BY LATE EVENING  
(50M): W'LY 10-14 SOON BACKING SSW INCREASING 43-48 BY  
LATE AFTERNOON,EASING 38-43 BY LATE EVENING  
SIG WAVE: 1.0-1.5 RISING AROUND 5.0 BY LATE EVENING  
MAX WAVE: 1.5-2.5 RISING AROUND 8.0 BY LATE EVENING  
WEATHER: PARTLY CLOUDY/CLOUDY BECOMING OVERCAST/MISTY  
WITH RAIN BY LATE AFTERNOON/EVENING  
5A.LONG RANGE OUTLOOK: SUN 08-Oct 2000, AND MON 09-Oct 2000,  
WIND(10M): SSW 30-34 EASING 16-20 BY MIDDAY SUNDAY THEN  
AGAIN INCREASING 38-42 BY EARLY MONDAY MORNING THEN  
VEERING SW DECREASING 20-24 BY LATE MON.EVENING  
SIG WAVE: AROUND 5.0 FALLING 2.5-3.0 BY MIDDAY SUNDAY THEN  
RISING 6.0-6.5 BY MID MORNING MONDAY,FALLIN AROUND  
4.0 BY LATE MONDAY EVENING

N.B. THE HIGHEST INDIVIDUAL WAVE THAT MAY BE EXPERIENCED IS OF THE  
ORDER OF TWICE THE SIGNIFICANT WAVE HEIGHT.

Figure 1: Complete wave and weather forecast issued 5 October 2000 a.m. (filename: 5Oct2000\_03.prm). Anonymised items in triple square brackets ([[...]]).

05-10-00

05/06	SSW	18	22	27	3.0	4.8	SSW	2.5	9
05/09	S	16	20	25	2.7	4.3	SSW	2.3	9
05/12	S	14	17	21	2.5	4.0	SSW	2.2	9
05/15	S	14	17	21	2.3	3.7	SSW	2.2	8
05/18	SSE	12	15	18	2.4	3.8	SSW	2.3	8
05/21	SSE	10	12	15	2.4	3.8	SSW	2.4	8
06/00	VAR	6	7	8	2.4	3.8	SSW	2.4	8
06/03	VAR	8	10	12	2.3	3.7	SSW	2.3	8
06/06	NNW	10	12	15	2.1	3.4	SSW	2.1	8
06/09	NNW	12	15	18	1.9	3.0	SSW	1.8	8
06/12	NNW	14	17	21	1.7	2.7	SSW	1.5	8
06/15	NW	14	17	21	1.6	2.6	SW	1.3	8
06/18	NW	12	15	18	1.4	2.2	WSW	1.3	9
06/21	NW	10	12	15	1.4	2.2	WSW	1.3	9
07/00	W	8	10	12	1.3	2.1	WSW	1.2	9
07/03	SW	12	15	18	1.3	2.1	WSW	1.2	10
07/06	SSW	16	20	25	1.5	2.4	WSW	1.1	10
07/09	SSW	24	30	37	1.8	2.9	WSW	0.9	10
07/12	SSW	30	37	46	2.7	4.3	SSW	0.6	10
07/15	SSW	34	42	53	3.4	5.4	SSW	0.4	11
07/18	SSW	36	45	56	4.0	6.4	SSW	0.6	11
07/21	SSW	34	42	52	4.5	7.2	SSW	0.8	11
08/00	SSW	32	40	50	5.0	8.0	SSW	1.2	11

Figure 2: Complete wind data file for 5 October 2000 (filename: 5Oct2000\_03.tab).

## 3.2 Trainable systems

data/outputs/system-generated/trainable-systems/

To create inputs to our generators, the input vectors as they appear in the corpus (see Section 2.2) are augmented with the following information in an automatic preprocessing phase: whether the change in wind direction compared to the preceding 7-tuple is clockwise or anti-clockwise; whether change in wind speed is an increase or a decrease; and whether a 7-tuple is the last in the vector.

Then, the augmented 7-tuples are converted into a system-specific representation (e.g. in the case of the PCFG systems, this is nonterminals with arguments). Because of the differences between system types, each takes slightly different inputs.

For details of the generator building, training methods and input representations for each type of system below, please refer to Belz & Kow, 2010 [3].

### 3.2.1 PCFG Systems

```
data/outputs/system-generated/trainable-systems/PCFG/fold*/{test,train}/*.pcfg-2gram.1
data/outputs/system-generated/trainable-systems/PCFG/fold*/{test,train}/*.pcfg-greedy.1
data/outputs/system-generated/trainable-systems/PCFG/fold*/{test,train}/*.pcfg-random.1
data/outputs/system-generated/trainable-systems/PCFG/fold*/{test,train}/*.pcfg-roulette.1
data/outputs/system-generated/trainable-systems/PCFG/fold*/{test,train}/*.pcfg-viterbi.1
```

We have included the outputs of five *p*CRU generators for the METEO domain created previously [1]. The *p*CRU base grammar for the METEO data is a set of generation rules with atomic arguments that convert an input into a set of NL forecasts.

A probability distribution over the base generator was obtained by the multi-treebanking method [1] from the Prodigy-METEO corpus. This method first parses the corpus with the base CFG and then obtains rule-application frequency counts from the parsed corpus which are used to obtain a probability distribution by straightforward maximum likelihood estimation. If there is more than one parse for a sentence then the frequency count increment is equally split over rules in alternative parses.

### 3.2.2 PSCFG Systems

```
data/outputs/system-generated/trainable-systems/PSCFG/fold*/{test,train}/*.pscfg-semantic.1
data/outputs/system-generated/trainable-systems/PSCFG/fold*/{test,train}/*.pscfg-unstructured.1
```

We created two probabilistic synchronous CFG (PSCFG) generators for the METEO domain using WASP<sup>-1</sup> [7]. The main task here was to create a CFG for wind data input representations. We used two different grammars (resulting in two dif-

ferent generators). The ‘unstructured’ grammar encodes raw corpus input vectors augmented as described in Section 2.2, whereas the ‘semantic’ grammar encodes representations with recursive predicate-argument structure that more resemble semantic forms. These were produced automatically from the raw input vectors.

Both the PSCFG-unstructured and the PSCFG-semantic generators were built in the same way, by feeding the CFG for wind data representations and the corpus of paired wind data representations and forecasts to WASP<sup>-1</sup> which then created PSCFGs from it.

### 3.2.3 PBSMT Systems

```
data/outputs/system-generated/trainable-systems/PBSMT/fold*/{test,train}/*.pbsmt-structured.1
data/outputs/system-generated/trainable-systems/PBSMT/fold*/{test,train}/*.pbsmt-unstructured.1
```

We also created four generators with the MOSES toolkit [4]. The main question here was how to represent the ‘source language’ inputs (the wind data). While SMT methods are often applied with no linguistic knowledge at all (and are therefore blind as to whether paired inputs and outputs are NL strings or something else), it was not clear how well they would cope with the task of mapping from number/symbol vectors to NL strings. We initially tested two different input representations [2], one of which was simply the augmented corpus input vectors as described above (PBSMT-unstructured), and another in which the individual 7-tuples of which the vectors are composed are explicitly marked by predicate-argument structure (PBSMT-structured). As in Wong & Mooney’s content-to-text generation work [7] we wanted to test the effect of treating the structure markers as tokens.

## 4 Recreating the complete 5-fold version of Prodigy-METEO

```
scripts/mkcorpus-5fold.pl
```

One of the main principles in designing this release of Prodigy-METEO was to avoid duplication of data files.

This means that in order to comparatively evaluate new systems against the existing systems for which outputs are included in this release, outputs for all systems have to be ‘aligned’ by creating either the 5-fold version of Prodigy-METEO (for which we are providing a script as described in this section), or some other aligned version.

For example, if you want to create a new trainable system and you want to compare it to our systems, you need to create the 5-fold version. Then for each fold, train your system on the \*.num.1/\*.prn.1 pairs in the `train/` subdirectory, and evaluate it on the dates in the `test/` subdirectory. Evaluation scores then need to be averaged over the 5 folds.

Running the `mkcorpus-5fold.pl` script creates the following new directory immediately under the top `Prodigy-METEO-beta/` top directory (here, only one date and all files for it are shown as an example):

```
Prodigy-METEO-5fold/
  fold1/
    test/
      ...
      3Apr2002_03.hyb.1
      3Apr2002_03.num.1
      3Apr2002_03.pbsmt-structured.1
      3Apr2002_03.pbsmt-unstructured.1
      3Apr2002_03.pcfg-2gram.1
      3Apr2002_03.pcfg-greedy.1
      3Apr2002_03.pcfg-random.1
      3Apr2002_03.pcfg-roulette.1
      3Apr2002_03.pcfg-viterbi.1
      3Apr2002_03.prn.1
      3Apr2002_03.pscfg-semantic.1
      3Apr2002_03.pscfg-unstructured.1
      ...
    train/
      ...
  fold2/
    ...
  fold3/
    ...
  fold4/
    ...
  fold5/
    ...
```

**NB:** In order to be able to compare directly with the SumTime-Hybrid system (for which we did not have outputs for every date), this process deletes a small number of dates (< 10) from each test set.

## 5 Example Use of Prodigy-METEO

1. Download `Prodigy-METEO_pre-alpha.zip` from [http://www.nltg.brighton.ac.uk/home/Anja.Belz/Prodigy-METEO\\_pre-alpha.zip](http://www.nltg.brighton.ac.uk/home/Anja.Belz/Prodigy-METEO_pre-alpha.zip)
2. Unzip `Prodigy-METEO_pre-alpha.zip`
3. Run `mkcorpus-5fold.pl`
4. Build your system and train it 5 times using only the data in the train subdirectory of one of the folds in `Prodigy-METEO-5fold`, each time.
5. Run each version of your system on the inputs in the test directory of the corresponding fold, so that you have added test data outputs for your systems for each date in `Prodigy-METEO-5fold/fold*/test/`
6. For each of the `Prodigy-METEO-5fold/fold*/test/` test sets, and each system you want to evaluate, run your chosen evaluation script(s), e.g. BLEU, then average scores over the 5 folds.

## 6 Contacts

Anja Belz, Brighton University, UK (a.s.belz@brighton.ac.uk)

Eric Kow, Brighton University, UK

Prodigy Website: <http://www.nltg.brighton.ac.uk/home/Anja.Belz/Prodigy>

## References

- [1] BELZ, A. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering* 14, 4 (2008), 431–455.
- [2] BELZ, A., AND KOW, E. System building cost vs. output quality in data-to-text generation. In *Proceedings of the 12th European Workshop on Natural Language Generation* (2009).
- [3] BELZ, A., AND KOW, E. Assessing the trade-off between system building cost and output quality in data-to-text generation. In *Empirical Methods in Natural Language Generation*, E. Kraemer and M. Theune, Eds. Springer, 2010. To appear.
- [4] KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A., AND HERBST, E. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)* (2007), pp. 177–180.
- [5] REITER, E., AND BELZ, A. An investigation into the validity of some metrics for automatically evaluating nlg systems. *Computational Linguistics* 35, 4 (2009).
- [6] SRIPADA, S., REITER, E., HUNTER, J., AND YU, J. SUMTIME-METEO: A parallel corpus of naturally occurring forecast texts and weather data. Tech. Rep. AUCS/TR0201, Computing Science Department, University of Aberdeen, 2002.
- [7] WONG, Y. W., AND MOONEY, R. Generation by inverting a semantic parser that uses statistical machine translation. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL'07)* (2007), pp. 172–179.