Chia-Yen Wu
chiayen@avaya.com
June 6, 2006

The MLB Language

1. Introduction

The MLB (Major League Baseball) language is designed to help an interested party determine certain characteristics of a baseball club. These characteristics are not always apparent, even if the party was in attendance. It will be useful in measuring basic statistics, as well as judging the behavior of an entire team.

1.1 Background

Before and after each individual MLB contest, a common set of resources is usually available to any user of the 'world wide web.' First, prior to the start of the game, the 'preview' gives the user lineup information – namely starting pitchers and batting orders. Post-game data includes a written 'recap', a 'box score', and 'play-by-play'. The recap provides a summary of the game but does not provide any actual data. The box score provides basic player data: hits, runs, home runs for the hitters, innings, strike outs and runs allowed for the pitcher.

Some baseball enthusiasts base the quality of a team or player simply based off the box score. For instance, the box score in *Figure 1* depicts a team that was able to obtain a barrage of hits and runs.

YANKEES (31-21)				
Lineup	**AB**	**R**	**H**	**RBI**
J. Damon CF	5	1	1	0
M. Cabrera LF	4	1	1	0
J. Giambi DH	5	1	2	1
A. Rodriquez 3B	3	1	0	0
J. Posada C	3	1	1	3
R. Cano 2B	5	1	4	0
A. Phillips 1B	4	0	3	2
T. Long RF	5	0	0	0
M. Cairo SS	4	0	0	0

Figure 1.

However, did this team score these runs early or late? Did they take advantage of every opportunity, or botched attempts to score even more runs? These details can be derived from the provided 'play-by-play'. The MLB language uses the play-by-play to calculate these non-tracked statistics.

2. Syntax

Of all the major sports websites that possess play-by-play data, 'cbs.sportsline.com' currently distinguishes itself from the rest because it also tracks pitch-by-pitch data. The MLB language syntax will exactly replicate that of the website's play-by-play, minus the extraneous HTML that controls the layout and text font.

2.1 At-Bat

Jason Giambi: Strike looking, **Giambi homered to right, Cairo and Cabrera scored.**

Alex Rodriguez: Ball, Ball, Strike looking, Strike (foul tip), Ball, **Rodriguez walked.**

Figure 2.

Figure 2 is a screenshot of two lines from the play-by-play of the June 5th contest between the Red Sox and the Yankees. Once again, the code that the MLB compiler and interpreter will use will not contain the table rows and columns, the bold text, or the colored font.

Each line of code must contain the following:
- The name of the player followed by a colon.
- The last name of the player repeated once or more after the colon.
- The line ends with a period.
- Pitches and other details are separated by commas.

Currently, MLB language does not track the location of each ball placed into play (e.g. to right, to shortstop).

2.2 Types of Hits

Once the last name of the player is repeated after the colon, the following keyword(s) will indicate whether the batter obtains a 'base hit', reaches base, or it commits an out.

Types of base hits are clearly defined: 'singled', 'doubled', 'tripled', and 'homered'. *
A player can reach base if the keyword(s) following the last name is: walked, intentionally walked, or hit by pitch.
A player will commit an out if any other keyword follows the last name.

* The MLB language will make two changes to the play-by-play. It will change 'reached on bunt single' and 'reached on infield single' to the keyword 'singled'. The coder should be aware of this change.

2.3 Inning Summary

End of Inning (3 Runs, 4 Hits, 1 Error)

Figure 3.

Figure 3 displays the end of inning summary that occurs after each set of 3 outs. This indicates to the interpreter than an inning has ended, and it should clear its knowledge of runs scored, hits, errors, and any man on base. The end of inning line of code is very consistent. It always contains the keywords 'End of Inning'. Followed by: the open parenthesis, NUM, space, "Runs", ",", space, NUM, space, "Hits", ",", space, NUM, space, "Error", close parenthesis. The 'close parenthesis' signifies the end of the line for the inning summary code snippet.

2.4 Pitches

A set of pitch data will be contained between the colon and the conclusion of an at-bat. Looking at *Figure 2*, you will notice Jason Giambi homered on the second pitch. Alex Rodriguez walked after six pitches.

The complete set of pitches are: Ball, Ball (pitchout), Intentional Ball, Strike looking, Strike swinging, and Strike (foul tip). A Pickoff Attempt is not considered a pitch.

Note that a player who makes contact with the ball on the final pitch of an at-bat, that pitch is considered a strike. All non-contact pitches are considered balls.

2.5 Comments

ROCKIES 8TH
Humberto Cota catching
Eli Marrero hit for Jose Mesa
Salomon Torres pitching:
Eli Marrero: Ball, Strike looking, Ball, **Marrero grounded out to shortstop**.
Jamey Carroll: Ball, Ball, Ball, Strike looking, Strike looking, **Carroll homered to left**.

Figure 4.

Substitutions and pitching changes displayed in *Figure 4* that do not follow the syntax of an at-bat or an inning summary are considered comments. However, the Team name followed by the inning number will be used by the MLB language for tracking game duration.

3. Output

The MLB interpreter will create statistics for each team on the following statistics: RoBA, NOPS, LIPS, RAAER, and PSPA.

3.1 RoBA

The 'Runners on Base Average' divides the number of base hits obtained by a team divided by the number of official at-bats while a player is on first, second, and/or third base. This can be compared with the overall batting average

of the team to determine if the ball club takes advantage of scoring opportunities. The range of this value is between .000 and 1.000. Unless a Perfect Game is thrown, this value is not NULL.

3.2 NOPS

The On-base Plus Slugging Percentage is a combined calculation of two common statistical categories, and popularized by the book/philosophy, Moneyball. The on-base percentage is calculated by summing the hits, walks, and hit by pitches, and dividing it by the number of plate appearances. The slugging percentage is calculated by dividing the total number of bases divided by at-bats. The 'New OPS' calculation by the MLB language will be adjusted and calculated by assigning the following values:
Single: 2.0
Double: 3.0
Triple: 4.0
Homerun: 5.0
Walk, Hit-by-pitch: 1.0
Out: 0
Strike-out: -1.0

These values will be totaled and divided by the number of plate appearances. The possible range is -1.0 to 5.0 but is generally between .500 and 1.000. A team with a high NOPS is getting on base and hitting the ball for power.

3.3 LIPS

The 'Late-Inning Pressure Situation' statistic for a team is calculated when the game is in the 7th inning or later, and the scores of the two teams differ by two runs or less. The MLB language will track the Batting Average of the team's LIPS by dividing the number of hits by the number of at-bats during this time period. The range of this value is between .000 and 1.000 but can also be NULL. A team that performs well in late innings demonstrates reliability during important circumstances.

3.4 RAAER

The 'Runs Allowed After Error Ratio' is evaluated by dividing the number of runs scored after an error occurs in the same inning. After an at-bat where an error occurs, the MLB interpreter will count the number of runs scored until the Inning Summary line of code is reached. The range of this value is 0.0 or more. It is NULL if no errors were committed during the game. The RAAER can help determine the frustration level of a team's pitcher when its defense falters. A low ratio can indicate that the team remains calm and focused.

3.5 PSPA

The 'Pitches Seen Per At-Bat' is calculated by averaging the number of pitches seen per at-bat for all hitters on the team. It counts all balls, strikes, fouls, and includes the final pitch that terminates the at-bat. The range of this value is 1.0 or more. A low value shows that the hitters are immature and rush at-bats; a high value can indicate that a team has good plate discipline.

4. Sample input/output

NATIONALS 7TH

Matt LeCroy hit for Shawn Hill

Horacio Ramirez pitching:

Matt LeCroy: Ball, Ball, Strike looking, Foul, **LeCroy struck out on foul tip.**

Alfonso Soriano: Strike looking, Ball, **Soriano grounded out to first.**

Royce Clayton: Ball, Ball, **Clayton grounded out to right.**

End of Inning (0 Runs, 0 Hits, 0 Errors)

BRAVES 7TH

Jon Rauch pitching:

Marcus Giles: Ball, Ball, Foul, Ball, **Giles fouled out to first.**

Edgar Renteria: Ball, Foul, Ball, Foul, **Renteria struck out swinging.**

Chipper Jones: Ball, Strike looking, Foul, Foul, Ball, **Jones grounded out to second.**

End of Inning (0 Runs, 0 Hits, 0 Errors)

NATIONALS 8TH

Horacio Ramirez pitching:

Jose Vidro: Strike looking, Strike looking, Ball, **Vidro struck out swinging.**

Nick Johnson: Ball, Ball, Foul, **Johnson singled softly to right.**

Ryan Zimmerman: Ball, **Zimmerman reached on fielder's choice to first, Johnson out at second.**

Marlon Byrd: Ball, Foul, Ball, Ball, Pickoff attempt to first, **Byrd flied out softly to center.**

End of Inning (0 Runs, 1 Hit, 0 Errors)

BRAVES 8TH

Jon Rauch pitching:

Andruw Jones: Strike looking, Ball, Strike looking, **Jones homered to left center.**

Bill Bray relieved Jon Rauch.

Adam A. LaRoche: **LaRoche safe at first on 2nd baseman Vidro's fielding error.**

Jeff Francoeur: Ball, Strike swinging, Foul, **Francoeur singled to center, LaRoche to second.**

Ryan Langerhans: Bunt foul, Bunt foul, Ball, **Langerhans struck out bunting foul.**

Brayan Pena: **Pena grounded into double play second to shortstop to first, Francoeur out at second.**

End of Inning (1 Run, 2 Hits, 1 Error)

NATIONALS 9TH

Kenny Ray pitching:

Robert Fick: Bunt foul, Strike looking, Ball, Ball, Foul, Foul, **Fick flied out to right**

Marlon Anderson hit for Damian Jackson.

Marlon Anderson: Ball, Strike looking, Strike looking, **Anderson struck out swinging**

Daryle Ward hit for Bill Bray.

Daryle Ward: Ball, Ball, Strike looking, Ball, Foul, **Ward struck out swinging.**

End of Inning (0 Runs, 0 Hits, 0 Errors)

Figure 5.

Figure 5 shows sample code (three total innings) for a contest between the Nationals and Braves. These three innings are considered a 'Late Inning Pressure Situation' because the Braves were leading 4-3 after 6 complete innings. The MLB interpreter will output the following for each team:

Nationals
RoBA: .000

NOPS: -.200
LIPS: .100
RAAER: 0.0
PSPA: 4.3

Braves
RoBA: .333
NOPS: .625
LIPS: .250
RAAER: null
PSPA: 3.75