How do you pronounce 07-1191?

Franklin Chen
FranklinChen@cmu.edu

Carnegie Mellon University

Pittsburgh Perl Workshop 2010
October 9, 2010

Goals of this talk

- Describe the specific problem to solve, in context.
- Describe a solution in Perl.
- Illustrate Perl's support for the three great programmer virtues:

Goals of this talk

- Describe the specific problem to solve, in context.
- Describe a solution in Perl.
- Illustrate Perl's support for the three great programmer virtues:
  - Laziness

Goals of this talk

- Describe the specific problem to solve, in context.
- Describe a solution in Perl.
- Illustrate Perl's support for the three great programmer virtues:
  - Laziness
  - Impatience

Goals of this talk

- Describe the specific problem to solve, in context.
- Describe a solution in Perl.
- Illustrate Perl's support for the three great programmer virtues:
  - Laziness
  - Impatience
  - Hubris

My work at CMU

- Research programmer at Carnegie Mellon University
- TalkBank project under Brian MacWhinney, Professor of Psychology and Modern Languages
- Text processing

My work at CMU

- Research programmer at Carnegie Mellon University
- TalkBank project under Brian MacWhinney, Professor of Psychology and Modern Languages
- Text processing
  - Parsing and validation of data in many formats
  - Converting between formats, especially to our CHAT
  - Our XML Schema (`http://talkbank.org/software/talkbank.xsd`)

Supreme Court (SCOTUS) oral argument transcript

Example excerpt:

Case 09-571 Argued this Wednesday (October 6, 2010), first week of 2010 term

Snippet
```
MR. COONEY: Your Honor, I think there are
two very quick answers to that. If one
looks at J.A. 550 to 551, which was
Mr. Glas, the grand jury prosecutor's,
testimony.
```

Conversion to CHAT format

Compare:

Original MR. COONEY: Your Honor, I think there are two very quick answers to that. If one looks at J.A. 550 to 551, which was Mr. Glas, the grand jury prosecutor's, testimony.

Conversion to CHAT format

Compare:

Original

MR. COONEY: Your Honor, I think there are two very quick answers to that. If one looks at J.A. 550 to 551, which was Mr. Glas, the grand jury prosecutor's, testimony.

CHAT

*COONE: Your Honor , I think there are two very quick answers to that .
*COONE: If one looks at <J_A> [= J.A.] <five fifty> [= 550] to <five fifty-one> [= 551] , which was <Mister> [= Mr.] Glas , the grand jury prosecutor's , testimony .

Validated XML equivalent to CHAT

```
<u who="COONE" uID="u667"><w>If</w><w>one</w>
<w>looks</w><w>at</w><g><w>J_A</w>
<ga type="explanation">J.A.</ga></g><g><w>five</w>
<w>fifty</w><ga type="explanation">550</ga></g>
<w>to</w><g><w>five</w><w>fifty-one</w>
<ga type="explanation">551</ga></g><s type="comma"/>
<w>which</w><w>was</w><g><w>Mister</w>
<ga type="explanation">Mr.</ga></g><w>Glas</w>
<s type="comma"/><w>the</w><w>grand</w><w>jury</w>
<w>prosecutor's</w><s type="comma"/>
<w>testimony</w><t type="p"></t></u>
```

CHAT format

- Fully defined, rigorously machine-parseable
- Analysis tools written to process CHAT, e.g., CLAN (http://childes.psy.cmu.edu/clan/)
- Want spoken representation
  - Link with audio and video clips, put online
  - Perform morphological and phonological analysis
- XML is just the next level for interoperability with other formats and tools

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	one tenth
decimal	20.4	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	ten o'clock
clock time	10:01	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	ten o'clock
clock time	10:01	ten oh one
ordinal	101st	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	ten o'clock
clock time	10:01	ten oh one
ordinal	101st	one hundred and first
year range	1941-42	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@I two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	ten o'clock
clock time	10:01	ten oh one
ordinal	101st	one hundred and first
year range	1941-42	nineteen forty-one to forty-two
plural	'20s	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@l two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	ten o'clock
clock time	10:01	ten oh one
ordinal	101st	one hundred and first
year range	1941-42	nineteen forty-one to forty-two
plural	'20s	twenties
case number	00-1011	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@I two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	ten o'clock
clock time	10:01	ten oh one
ordinal	101st	one hundred and first
year range	1941-42	nineteen forty-one to forty-two
plural	'20s	twenties
case number	00-1011	oh oh ten eleven
case number	00-11291	

Potpourri of conversions to CHAT

Category	Original	CHAT
phone number	911	nine one one
section with Roman	(1)(A)(ii)	one A@I two
fraction	1/10	one tenth
decimal	20.4	twenty point four
decimal	.0001	point oh oh oh one
clock time	10:00	ten o'clock
clock time	10:01	ten oh one
ordinal	101st	one hundred and first
year range	1941-42	nineteen forty-one to forty-two
plural	'20s	twenties
case number	00-1011	oh oh ten eleven
case number	00-11291	oh oh eleven two ninety-one

Stealing other people's code

The Comprehensive Perl Archive Network (`CPAN`) is the lazy Perl programmer's friend.

Stealing other people's code

The Comprehensive Perl Archive Network (CPAN) is the lazy Perl programmer's friend.

- `use Lingua::EN::Numbers qw(num2en num2en_ordinal);`

Stealing other people's code

The Comprehensive Perl Archive Network (CPAN) is the lazy Perl programmer's friend.

- `use Lingua::EN::Numbers qw(num2en num2en_ordinal);`
- `use Lingua::EN::Inflect qw(PL);`

Stealing other people's code

The Comprehensive Perl Archive Network (CPAN) is the lazy Perl programmer's friend.

- `use Lingua::EN::Numbers qw(num2en num2en_ordinal);`
- `use Lingua::EN::Inflect qw(PL);`
- `use Text::Roman;`

Stealing other people's code

The Comprehensive Perl Archive Network (CPAN) is the lazy Perl programmer's friend.

- `use Lingua::EN::Numbers qw(num2en num2en_ordinal);`
- `use Lingua::EN::Inflect qw(PL);`
- `use Text::Roman;`
- `Dist::Zilla`

Integrating other people's code

Main workhorse: a recursive function `happeningToWords` that takes a string and returns a list of words.

- Use regular expressions to pick string apart
- Use CPAN libraries on small enough units

Example: plurals

```
# Plural of something looking numeric at the end.
elsif ($t =~ /^
                (
                .+
                \d
                )
                s
                $/x) {
  my @words = happeningToWords($1);
  $words[$#words] = PL($words[$#words]);
  return @words;
}
```

Example: ordinals

```
# Take apart ordinals: 21st 22nd 23rd 137,534th
elsif ($t =~ /^
            (
            [\d\,]+
            )
            (?:
            st
            |nd
            |rd
            |th
           )
            $/x) {
  return num2en_ordinal($1);
```

Example: case numbers

```
# dd-ddd
elsif ($t =~ /^
              (\d\d)
              -
              (\d\d\d)
              $/x) {
  return (case2ToWords($1), case3ToWords($2));
}
```

Laziness in testing

Using `Test::More`, `Test::Exception` is really easy.

Expected output

```
    [
      '1:00',
      'one o\'clock',
      'clock time'
    ],
```

Error to catch

```
    [
      '0and',
      'spurious numeric character'
    ],
```

Impatience in testing

Impatience with the computer churning out poor output leads me to add more test cases and let the computer help me.

- 138 test cases, both positive and negative
- A lot of `TODO` cases

Impatience in testing

Impatience with the computer churning out poor output leads me to add more test cases and let the computer help me.

- 138 test cases, both positive and negative
- A lot of `TODO` cases
  - Not yet implemented although possible
  - No unambiguously correct output (00 by Rehnquist or Roberts)

Complete distribution

- Code has 45 `elsif`, 12 `croak`
- `Test::Pod::Coverage` through `Dist::Zilla`
- Doing something despite odds and imperfection
  - Ambiguities
  - Heuristics imperfect; maybe machine learning helpful
- Packaging `Lingua::EN::Numbers::SCOTUS` for wider use

Conclusion

- Needed to convert written typography to spoken representation
- Perl did the job through Laziness, Impatience, Hubris
- Thank you to the Perl community for sharing!

Links for this talk

Slides for this talk

`http://franklinchen.com/ppw2010/slides.pdf`

Code for `Lingua::EN::Numbers::SCOTUS`

`http://franklinchen.com/ppw2010/Lingua-EN-Numbers-SCOTUS-1.0.tar.gz`

SCOTUS transcripts

Web sites that provide different formats and interactive audio-linked:

Definition of our CHAT format `http://childes.psy.cmu.edu/manuals/chat.pdf`

SCOTUS transcripts in our CHAT format with linked audio

`http://talkbank.org/browser/index.php?url=Meeting/SCOTUS/`

Oyez project: our collaborator's linked audio interface

`http://www.oyez.org/`

Official Supreme Court oral argument transcripts in PDF

`http://www.supremecourt.gov/oral_arguments/argument_transcripts.aspx`