It must be said that long before there were Processing sketches and Flash applets, people were exploring these sorts of typographical constructions in mass media and in fine art (Figure 3-14); we have been probing the boundary between letters as forms and letters as signs for a long time (Figure 3-15). The goal of these algorithmic explorations is to allow the wit and elegance of such examples to influence the representation of textual data.

Given this rather brief tour of the technical and aesthetic environment in which Wordle evolved, we’re now ready to look at Wordle’s technical and aesthetic choices in a bit more detail.

Figure 3-14. Herb Lubalin and Lou Dorfsman’s Typographicalassemblage (courtesy of the Center for Design Study)

Figure 3-15. Before we made pictures with words, we made words with pictures
How Wordle Works

Wordle is implemented as a Java applet. Some of the technical details I provide here will be in terms of Java-specific features. Nothing described here is impossible in other languages, using other libraries and frameworks, but Java’s strong support for Unicode text processing and 2D graphics (via the Java2D API) makes these things pretty straightforward.

Text Analysis

We’ll now take a step back and consider some of the fundamental assumptions that determine Wordle’s character. In particular, we have to examine what “text” is, as far as Wordle is concerned.

While this kind of text analysis is crude compared to what’s required for some natural-language processing, it can still be tedious to implement. If you work in Java, you might find my cue.language library useful for the kinds of tasks described in this section. It’s small enough, it’s fast enough, and thousands use it each day as part of Wordle.

Remember that natural-language analysis is as much craft as science,† and even given state-of-the-art computational tools, you have to apply judgment and taste.

Finding words

Wordle is in the business of drawing pictures of words, each having some weight, which determines its size. What does Wordle consider to be a “word”?

Wordle builds a regular expression (regex) that recognizes what it considers to be words in a variety of scripts, and then iteratively applies that regex to the given text, as illustrated in Example 3-1. The result is a list of words.

Example 3-1. How to recognize “words”

```java
private static final String LETTER = "[@+\p{javaLetter}\p{javaDigit}]";
private static final String JOINER = "[-.:/''\p{M}\u2032\u00A0\u200C\u200D~]";

/*
A word is:
   one or more "letters" followed by
   zero or more sections of
      one or more "joiners" followed by one or more "letters"
*/
private static final Pattern WORD =
    Pattern.compile(LETTER + "+\(\ +JOINER\ +\ +LETTER\ +\ +\)\+\)\+\);
```

* See http://github.com/vcl/cue.language.
† For an illuminating demonstration of this craft, see Peter Norvig’s chapter on natural-language processing in the sister O’Reilly book Beautiful Data.